Bayesian Optimal Design for Changepoint Problems

Juli Atherton

Department of Mathematics and Statistics,
McGill University, Montréal
Québec, Canada

February, 2007

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements of the degree of
Doctor of Philosophy

Copyright © Juli Atherton, 2007

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-32142-3
Our file Notre référence
ISBN: 978-0-494-32142-3

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Blackwell: "Within that area [Bayesian Statistics] it seems to me that one promising direction which hasn't been explored at all is Bayesian experimental Design."

DeGroot: "I think the reason there hasn't been so much done is because the problems are so hard."

Taken from a 1984 interview of David Blackwell by Moris DeGroot (published in Statistical Science, 1986)

Abstract

We consider optimal design for changepoint problems with particular attention paid to situations where the only possible change is in the mean. Optimal design for changepoint problems has only been addressed in an unpublished doctoral thesis, and in only one journal article, which was in a frequentist setting. The simplest situation we consider is that of a stochastic process that may undergo a change at an unknown instant in some interval. The experimenter can take n measurements and is faced with one or more of the following optimal design problems: Where should these n observations be taken in order to best test for a change somewhere in the interval? Where should the observations be taken in order to best test for a change in a specified subinterval? Assuming that a change will take place, where should the observations be taken so that that one may best estimate the before-change mean as well as the after-change mean? We take a Bayesian approach, with a risk based on squared error loss, as a design criterion function for estimation, and a risk based on generalized 0-1 loss, for testing. We also use the Spezzaferri design criterion function for model discrimination, as an alternative criterion function for testing. By insisting that all observations are at least a minimum distance apart in order to ensure rough independence, we find the optimal design for all three problems. We ascertain the optimal designs by writing the design criterion functions as functions of the design measure, rather than of the designs themselves. We then use the geometric form of the design measure space and the concavity of the criterion function to find the optimal design measure. There is a straightforward correspondence between the set of design measures and the set of designs. Our approach is similar in spirit, although rather different in detail, from that introduced by Kiefer. In addition, we consider design for estimation of the changepoint itself, and optimal designs for the multipath changepoint problem. We demonstrate why the former problem most likely has a prior-dependent solution while the latter problems, in their most general settings, are complicated by the lack of concavity of the design criterion function.

Résumé

Nous considérons, dans cette dissertation, les plans d'expérience bayésiens optimaux pour les problèmes de point de rupture avec changement d'espérance. Un cas de point de rupture avec changement d'espérance à une seule trajectoire se présente lorsqu'une séquence de données est prélevée le long d'un axe temporelle (ou son équivalent) et que leur espérance change de valeur. Ce changement, s'il survient, se produit à un endroit sur l'axe inconnu de l'expérimentateur. Cet endroit est appelé “point de rupture”. Le fait que la position du point de rupture soit inconnue rend les tests et l'inférence difficiles dans les situations de point de rupture à une seule trajectoire.

L'exploration d'un problème de point de rupture à une seule trajectoire s'accomplit souvent par le truchement des questions suivantes : y a-t-il eu changement? où le changement s'est-il produit? quelle était l'ampleur du changement? L'analyse de cas de point de rupture à une trajectoire s'effectue fréquemment de manière rétrospective, après une collecte de données faite uniformément sur l'intervalle temporel d'intérêt. Lors de ces analyses rétrospectives, le modèle est construit de telle manière que la loi *a priori* du point de rupture permet à la rupture de se produire uniquement là où une donnée a été relevée, et ce habituellement pour des raisons de commodité computationnelle. Puisque nous nous intéressons aux plans optimaux, nous choisissons plutôt, et de manière plus réaliste, une loi *a priori* qui permet à la rupture de se produire á tout endroit dans l'intervalle temporel d'intérêt. La première conséquence de cette modification est que nous ne pouvons plus obtenir la loi *a posteriori* du point

de rupture, et que nous ne pouvons donc plus estimer sa position. Une deuxième conséquence est que, pour obtenir des données conditionnellement indépendantes, nous devons exiger que les mesures soient prises à une distance minimale les unes des autres dans notre plan d'expérience.

Dans cette dissertation, nous considérons des plans bayésiens optimaux pour les tests de changement d'espérance, les tests de changement d'espérance dans un intervalle donné, et l'estimation des espérances avant et après le point de rupture. Nous ne sommes pas en mesure de considérer les plans bayésiens optimaux pour estimer la position du point de rupture car nos données ne peuvent actualiser la loi *a priori* du point de rupture. Malgré cela, nous construisons un plan d'expérience que nous croyons utile pour qui désire tirer une inférence concernant le point de rupture.

Nos plans optimaux résultent de la minimisation de fonctions-critères spécifiques. Nous minimisons le risque bayésien 0-1 généralisé afin d'obtenir des plans optimaux pour les tests d'hypothèses, ainsi que le risque bayésien quadratique pour obtenir des plans optimaux pour l'estimation. Nous utilisons de plus la fonction-critère de Spezzaferri, initialement conçue pour la sélection de modèle, pour concevoir des modèles optimaux pour les tests. Nos modèles optimaux pour tests et estimation se retrouvent tous dans un petit sous-ensemble de modèles qui placent les observations aussi près que possible des extrémités de l'intervalle d'intérêt.

Au centre de nos résultats se trouve une mesure de plan que nous avons construite et qui rappelle la mesure utilisée par Kiefer pour sa théorie d'approximation continue. L'optimisation convexe joue aussi un rôle important dans l'obtention de nos résultats. Bien que nous ne considérions que trois fonctions-critères (risque bayésien 0-1 généralisé, risque bayésien quadratique, et fonction-critère de Spezzaferri), les plans optimaux résultants s'appliquent à toute fonction-critère qui soit une fonction concave de notre mesure de plan.

Nous concluons en considérant brièvement le problème de point de rupture à

plusieurs trajectoires. Dans cette situation, plusieurs séquences de données sont relevées, chacune assortie d'espérances aléatoires avant et après le point de rupture. Lorsque les séquences partagent le même point de rupture, le problème à plusieurs trajectoires se réduit à un problème à une seule trajectoire. Lorsque chaque trajectoire possède son propre point de rupture, la détermination de plans bayésiens optimaux devient beaucoup plus complexe.

Acknowledgments

I thank my thesis advisor, David Wolfson, who has played a huge role, not just in the writing of this thesis, but also in all aspects of my graduate career. David has always made time for me, often at quite short notice, and has always made me feel welcome. I have learned a tremendous amount during the many enjoyable discussions we have had regarding this thesis and statistics in general. It was David who first introduced me to the topic of Bayesian optimal design for changepoint problems; he also suggested many of the specific design problems considered in this thesis. I thank him, as well, for a very detailed revision of my work.

I also thank my co-advisor Alain Vandal. This work improved substantially with Alain's editing. Moreover, I benefited greatly from his mathematical statistics courses. In particular, there I first learned about natural exponential families and decision theory; both of these topics appear in the current thesis. I am also very grateful to him for translating my abstract into French.

Another person who played a very large role in this thesis is my friend and co-author Benoit Charbonneau. It was with Benoit's guidance that I was able to extend the log-concave result of Chapter 5 to high dimensions. Benoit also spent a great deal of time proof reading and pointing out any omissions and mistakes. In addition, he helped me improve the lucidity of this work; a major milestone was the day he helped me to simplify my notation, which initially was very cumbersome!

This work follows on from the doctoral thesis of Xiaojie Zhou (1997), which was carried out under the supervision of David Wolfson and Lawrence Joseph. I thank Xiaojie and Lawrence for welcoming me to their project, and there is no doubt that their influence is felt in this thesis. I also thank Lawrence for the discussions we have had. Other people I have spoken to include Phillippe Poulin, Charles Fortin, Paul Tupper, Dave Stephens, Adrian Vetta, Russell Steele, Vanamamalai Seshadri, Stephanie Atherton, Jose Correa, Gordon Craig and Bruce Shepherd. I thank all of these people for taking the time to listen to me and think about my inquiries. Thanks also to Olivier Dubois for making the latex style files for his thesis available over the web.

I have been surrounded by wonderful people in the Mathematics and Statistics Department here at McGill. I thank all of the administrative staff on the 10th floor, in particular Carmen Baldonado, the graduate student secretary, Greg Lebaron, the computer administrator, and, Christina Girardi, who has been my gateway to David for the last couple of years. In addition to my advisors, I also thank the professors in the department, especially those that I have taken classes from, been a teaching assistant for, or taught under. I especially thank Russell Steele and Keith Worsley. My fellow classmates have made my time at McGill more interesting and enjoyable. Specifically, thanks to Geva Maimon for all the soccer games, to Liz Turner for the journal club, and to Geneviève Lefebvre for the many study breaks.

Lastly, I thank my family. I thank my parents in Newfoundland and my sister and her husband in Utah for their support and good company. I also thank them for having a last-minute look at my thesis for typos. I especially thank my partner Adrian Vetta who, as well as talking to me from time to time about my work, has helped with the preparation of this document. Thanks also to Adrian and his family for their support. In particular thanks to Amber and Elysia who have entertained Adrian and me with riveting phone conversations, storybooks, and games.

My doctoral studies were supported, in part, by grants from the Institut des

Sciences Mathématiques and the Natural Science and Engineering Research Council of Canada.

Table of Contents

Abstract iii

Résumé v

Acknowledgments ix

Statement of Originality 1

1 Introduction 3

1.1 Motivation and Description of Thesis 3

1.2 Literature Review 7

1.2.1 Changepoint Problems 8

1.2.2 Optimal Design 10

1.3 Optimal Design in this Thesis 14

1.3.1 Outline of Thesis 15

2 Design Criterion Functions 19

2.1 Decision Theory 20

2.2 Bayes Risks as Design Criterion Functions 23

2.3 Scoring Rule Utilities 24

2.4 The Spezzaferri Criterion 26

2.5 Comments on the Spezzaferri Criterion 27

3 Preliminaries **31**

3.1 Convex Optimization 31

3.2 Differential Geometry 39

3.3 Exponential Families 46

3.3.1 Natural Exponential Families 47

3.3.2 General Exponential Families 48

3.3.3 Conjugate Priors for NEFs 49

4 The Design Measure **53**

4.1 The Single-Path Changepoint Model 53

4.2 The Design Space 54

4.3 The Design Measure 56

4.4 The Mapping from the Design x to the Design Measure π 58

5 The Shape of $G_f(\mathbb{X}^n)$ **61**

5.1 Examples and Motivation 62

5.2 Proof for f Everywhere Greater than Zero 66

5.2.1 Part I: Affine Independence of $G_f(V)$ 66

5.2.2 Part II: Condition on f so that $G_f(\mathbb{X}^n) \subset \triangle_f$ 67

5.3 When f Equals Zero Over Part of $[0, T]$ 73

5.3.1 Type 1 Prior 75

5.3.2 Type 2 Prior 76

5.3.3 Type 3 Prior 76

5.4 When f has Positive Probability of no Changepoint 78

5.5 Optimal Designs for Design Criterion Functions Concave in π 79

5.6 Restriction to a Subset of Designs 80

5.6.1 Fixing x_n at Time T 80

5.6.2 Fixing x_q and x_{q+1} at Times t_1 and t_2 80

6 Optimal Designs for the Single-Path Changepoint Problem 83

6.1 Model and Dependence on π 84

6.2 Optimal Designs for Testing 86

6.2.1 Optimal Design for Testing for a Change 86

6.2.2 Optimal Design for Testing for a Change in a Subinterval . . . 89

6.3 Designs and Optimal Designs for Estimation 92

6.3.1 Optimal Designs for Estimating the Before-and-After-Change Means 92

6.3.2 Design Criterion Function for Estimation the Changepoint Location 102

6.4 Combining Criterion Functions 103

7 Particular Single-Path Changepoint Models 105

7.1 The NEF Single-Path Changepoint Problem 107

7.1.1 The Bayes Risk Based on Squared Error Loss 110

7.1.2 Simulations 114

7.2 The Common Changepoint Multi-Path Problem 124

7.2.1 The Model 125

7.2.2 The Bayes Risk Based on Squared Error Loss 128

7.2.3 Simulations 129

8 Optimal Design for the Multiple Changepoints Multi-Path Problem137

8.1 The Model 139

8.2 Estimation of Proportions 141

8.3 Estimation of the Hierarchical Before-and-After-Change Means 142

8.4 Simulations 144

9 Summary and Future Work 149

9.1 Thesis Summary 149

9.2 Future Work 150

9.2.1 Extensions to the Single-Path Optimal Design Problem 150

9.2.2 The Multiple Changepoints Multi-Path Problem 152

9.2.3 The Biphasic Regression Problem 152

9.2.4 The Cross-Sectional Data Changepoint Problem 153

A Remarks for Chapter 5 **155**

A.1 Remarks for Theorem 5.6 155

B Algebraic Details for Chapter 7 **157**

B.1 The Density $f(y|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k)$ 157

B.2 The Density $f(\bar{y}|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k)$ 159

B.3 Posterior Means and Variances of $\bar{\mu}_1$ and $\bar{\mu}_2$ 160

B.4 The Density $f(\bar{y}|\tau_x = k)$ 161

C Algebraic Details for Chapter 8 **163**

C.1 The Density $f(y|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k)$ 163

C.2 Posterior Means and Variances of $\bar{\mu}_1$ and $\bar{\mu}_2$ 165

C.3 The Density $f(y|\tau_x = k)$ 166

Statement of Originality

The idea of optimal design for changepoint problems is not new and has been previously addressed by Zhou (1997) and Bischoff and Miller (2000). This thesis corrects and considerably extends Zhou's formulation, proofs and results which are in a Bayesian setting. Bischoff and Miller (2000) discusses optimal design for changepoint problems in a frequentist setting, but assumes a known changepoint, in contrast to our setting which is based on an unknown changepoint.

Below we list the results and observations that are, to our knowledge, new. Some of the results listed as "new," perhaps do not merit this strict designation as they follow as direct easy consequences of more other substantive new results. We include them, nevertheless, for completeness.

Chapter 1: None.

Chapter 2: Theorem 2.7 and Theorem 2.8.

Chapter 3: None.

Chapter 4: Idea of specific design measure is new. Section 4.3 and Section 4.4.

Chapter 5: Lemma 5.1, Theorem 5.1, Lemma 5.2, Theorem 5.2, Lemma 5.3, Theorem 5.3, Lemma 5.4, Theorem 5.4, Theorem 5.5, Theorem 5.6, Theorem 5.7, and Theorem 5.8.

Chapter 6: Lemma 6.1, Lemma 6.2, Theorem 6.1, Lemma 6.3, Lemma 6.4, Theorem 6.2, Lemma 6.5, Lemma 6.6, Lemma 6.8, Theorem 6.3, Theorem 6.4, Theorem 6.5, and Theorem 6.6.

Chapter 7: Theorem 7.1, Lemma 7.1, Lemma 7.2, Theorem 7.2, and Theorem 7.3.

Chapter 8: Various new observations.

Chapter 9: None.

Chapter 1

Introduction

1.1 Motivation and Description of Thesis

This thesis considers Bayesian optimal designs for change-in-mean changepoint problems. In the single-path changepoint problem a sequence of data are collected along some time axis or equivalent. Initially the data are distributed about some mean and then immediately after some point, called the *changepoint*, the mean changes value. The unknown location of the changepoint and the fact that it is not directly observed are what makes testing and estimation for this problem difficult. In some cases it is not even clear if a change has occurred.

Traditionally, analyses of such changepoint problems have been done retrospectively on data collected at regular intervals throughout a period or distance of interest. It is usually assumed that the change can only occur at locations where data have been collected. Therefore, one has n observations $y = (y_1, \ldots, y_n)$ collected at n equally spaced locations. If r denotes the index location of the changepoint then the data $(y_1, \ldots, y_r)$ are distributed about the first value of the mean and the data $(y_{r+1}, \ldots, y_n)$ are distributed about the second value of the mean. Given the before-and-after-change means and the changepoint, the data are usually considered to be conditionally independent. The event $\{r = n\}$ is equivalent to no changepoint occurring.

Inference for such problems has been carried out in frequentist, Bayesian, and nonparametric settings. Interest lies in one or more of the following: testing for a change, estimating the location of the changepoint, and estimating the before-and-after-change mean values. Here we consider Bayesian optimal design for each of the three types of problem described above.

Since we are considering optimal designs, our interest lies in situations where it is impossible, or too expensive, to collect data throughout the period of interest. Also since the time or distance axes along which the data are usually collected are in reality continuous, we consider the situation where n measurements are collected in the *observation interval* $[0, T]$. Our goal is to determine where to place the n measurements in order to obtain the "best" inference possible. The only constraint we impose on our designs is that we insist all measurements are a minimum distance d apart. We do so to ensure, as is usually the case in changepoint analysis, that the observations are conditionally independent given the before-and-after-change means and the changepoint.

Importantly too, since we are considering an infinite number of possible designs in a continuous period for which many designs do not correspond to observations taken at regular intervals, it is unreasonable to insist that the change only occurs at locations where data are collected. Hence, in our model we allow the changepoint, denoted by τ, to occur at any point in the interval $[0, T]$. Here, the event $\{\tau = T\}$ is equivalent to no change.

Change-in-mean changepoint models are useful when making inference about an underlying stochastic process where the realized paths are essentially "horizontal" initially, then increase or decrease quickly, and are then essentially "horizontal" again. We do not assume continuous observation of such a stochastic process; rather, we model the joint distribution of n data points collected a minimum distance d apart. These changepoint models are often sufficient and convenient for making inference. Changepoint analyses are justified when, for example, an underlying process reaches a threshold that causes the process under study to change from one state to another (see Beckage et al. (2006)). See also Joseph et al. (1997) [p. 691] for further discussion about using a change-in-mean changepoint model for data that in fact display a gradual change.

The illustrative example below, from medicine, makes the ideas presented above concrete!

Example 1.1. Change in Mean Blood Pressure

Consider a patient to be treated for high blood pressure at some time point in $[0, T]$. *Once the data are collected three questions of interest might be: Did the treatment have an effect? When did the treatment take effect? What was the magnitude of the effect of the treatment? In the same spirit as Joseph et al. (1996), we make two assumptions when modelling data from the stochastic process that describes the changing blood pressure over time both before and after the treatment.*

The first is that the blood pressures form a Gaussian process. Conditional on the changepoint τ *we assume that there is a covariance stationary process before* τ *and a possibly different covariance stationary process after* τ. *We emphasize that* τ, *the time at which the treatment will take effect, might not coincide with the time at which the treatment will be administered, and that it is not known.*

The second assumption is that given the before-and-after-change blood pressure means and the changepoint, the random variables y_{t_1} *and* y_{t_2}, *representing the blood pressure at times* t_1 *and* t_2, *where* $t_1 < t_2$, *are roughly conditionally independent provided that* t_1 *and* t_2 *are separated by some sufficiently large distance* d.

If the treatment takes effect quickly, the first assumption that the mean changes abruptly is reasonable. The second assumption is needed since there does not exist a *continuous parameter stochastic process with all observations independent and having a finite variance; we need independence to construct our likelihood.*

Our goal is to answer one or more of the three questions posed at the start of this example. To do this we shall collect n blood pressure readings in the interval $[0, T]$. *The optimal designs developed in this thesis will help us answer the three questions as efficiently as possible.*

Since only one sequence of measurements (on a single patient) was taken, the problem described above is termed a single-path changepoint problem. One of our main results for the single-path problem is that, when testing for a changepoint or when estimating the before-and-after-change means, the optimal design is one of a relatively small set of designs placing observations as far as possible towards the ends of the observation interval; determining *the* optimal design thus becomes numerically feasible. Next, although we cannot directly estimate the changepoint since our data do not update the prior density for τ, we suggest designs to help make inference about the location of the change. In particular, we find designs that are optimal for testing if the change occurs in a subinterval $[t_1, t_2]$ of the observation interval $[0, T]$. Our results apply to any distribution for the data y and there is no requirement that the prior distributions (in our Bayesian setting) for the before-and-after-change means, be conjugate.

In this thesis we also consider Bayesian optimal designs for multi-path changepoint problems. In a multi-path problem, multiple sequences of measurements are collected. An example of a multi-path changepoint problem in the same setting as Example 1.1 is a clinical trial where the same treatment will be administered on many patients and n observations are to be taken on each patient. In fact, the setting just described is the situation in the paper by Joseph et al. (1996) who were concerned with making inference from data already collected, rather than with optimal design for data about to be collected. In our setting each patient is assumed to have a random before-change mean effect and a random after-change mean effect. The random before-and-after-change mean effects for all subjects are assumed to be distributed about hierarchical before-and-after-change means.

There are two types of multi-path problems, the common changepoint multi-path problem and the multiple changepoints multi-path problem. In the first type we insist that all subjects change at the same time; that is, they have a common changepoint τ. In this thesis, we show that the common changepoint multi-path problem simply reduces to a single-path problem. Therefore, all our optimal design results for the single-path problem apply to the common changepoint multi-path problem. More realistically, in the multiple changepoints multi-path problem, we allow each subject to have their own changepoint. For the multiple changepoints multi-path problem we consider design criterion functions for estimating the before-and-after-change hierarchical means, estimating the proportion of people who do not change, and estimating the proportion of people who change in a specific interval $[t_1, t_2]$. Finding the optimal design in the multiple changepoints multi-path problem is much more complicated than in the single-path problem.

1.2 Literature Review

Bischoff and Miller (2000) present an asymptotic frequentist optimal design for a biphasic regression when the location of the possible changepoint is known. Their designs are optimal for testing whether or not the change occurred. As in our Bayesian setting, their frequentist optimal design is to place observations at the ends of the interval. There are two important differences between their setting and ours: first, the location of the possible change is unknown in our model. Second, they allow design points to be coincident whereas we insist that the design points be a minimum distance d apart. This last requirement simplifies the likelihood by forcing conditional independence, while considerably increasing the difficulty of finding the optimal design. This thesis follows the Ph.D. thesis of Zhou (1997), which considered various Bayesian optimal design problems for single-path changepoint problems.

Since, to our knowledge, there is no other work in the specific area of optimal design for changepoint problems, we continue our review by providing an overview of changepoint problems in Section 1.2.1 and then providing an overview of optimal design in Section 1.2.2. Each area has a vast literature and hence we emphasize only results most relevant to this thesis.

1.2.1 Changepoint Problems

In this thesis, we address single changepoint problems with fixed sample size. There is, however, an extensive literature on sequential changepoint analysis, for example the paper by Carter and Blight (1981). Other variants of the basic change-in-mean problem include work by Krolewski et al. (1995) on changepoints in regression, by Christensen and Rudemo (1998) on multiple changepoints in a sequence of measurements, by Picard (1985) on changepoints in time series, and by Müller and Wang (1994) on changepoints in hazard functions. Changepoint problems appear in many settings. For instance, the aforementioned references apply to detecting ovulation in women, the study of diabetes, the study of disease in pigs, quality control, and survival analysis.

Our review concentrates specifically on non-sequential changepoint problems for which there is a possible change in the mean only. The optimal designs presented in this thesis are for parametric models of the type discussed by Henderson and Matthews (1993) to study the incidence of haemolytic uremic syndrome (HUS), and by Chu and Zhao (2004) to study cyclone activity.

Although changepoint problems were first introduced by Shewhart (1931), the subject had its formal start with the three seminal papers by Page: Page (1954) addressed sequential and non-sequential inspection schemes; Page (1955) used cumulative sums for a one sided test to detect a change; and Page (1957) extended the work in the second paper to a two-sided test. These articles by Page were the start of what has become an extensive literature on testing and estimation. The literature covers parametric and nonparametric problems in both the frequentist and Bayesian settings. In the frequentist parametric setting, maximum likelihood estimators (MLE) for estimating the changepoint location were developed by Hinkely (1970) for normally distributed data and by Hinkley and Hinkley (1970) for data with a binomial distribution. Hawkins (1977) considered the problem of testing for a change and Worsley (1986) included confidence intervals for the changepoint. In the nonparametric frequentist setting, both Bhattacharya and Johnson (1968) and Pettitt (1979) developed tests for a change. Chernoff and Zacks (1964) were the first to address the Bayesian parametric setting by estimating the current mean of a sequence of random variables and Smith (1975) introduced the type of single-path hierarchical Bayesian model used in this thesis. Carlin et al. (1992) proposed the use of Gibbs sampling in Bayesian heirarchical changepoint models. The Bayesian nonparametric setting is addressed in Muliere and Scarsini (1985) and Mira and Petrone (1994). Zacks (1983) reviews much of the earlier work mentioned above.

Common distributions occurring in changepoint problems are binomial for binary data, normal for continuous data, exponential for interval data, and Poisson for count data. As we will see in Section 3.3, all these distributions are natural exponential families (NEFs). Hence it is not surprising that researchers have considered all these distributions at once by studying changepoint problems for NEFs. Worsley (1986) considered the problem of testing for a change and estimating the changepoint location for an exponential family. He also discussed interval estimation for the changepoint. Ghorbanzadeh and Lounes (2001) carried out a Bayesian analysis to detect a change in an exponential family. More recently, Wu (2005) has written a book about CUSUM tests for changepoint problems with emphasis on exponential families.

The book by Chen and Gupta (2000) provides a good introduction to change-in-mean, change-in-variance, and change-in-regression-slope problems. They discuss parametric models in both the frequentist and Bayesian settings. A second book by Csörgö and Horváth (1998) addresses limit theorems for change-in-mean, change-in-variance, and change-in-regression-slope problems in both the parametric and non-parametric settings.

The multi-path changepoint problem was first introduced in the Ph.D. thesis by Joseph (1990), and later presented by Joseph and Wolfson (1992) and Joseph and Wolfson (1993). The change-in-mean multi-path problem has been used in the Bayesian setting for the analysis of blood pressure data of Lyle et al. (1987) (see Joseph et al. (1996)). The multi-path problem has also been extended to change-in-regression-slope models (see Joseph et al. (1999)). This extension was applied on measurements of cognitive decline in patients with Alzheimer's disease. Chu et al. (2005), applied the change-in-regression-slope multi-path model to CD4 cell counts in AIDS patients.

1.2.2 Optimal Design

The subject of optimal design lies in the area of experimental design. It involves determining a suitable design criterion function, and then minimizing or maximizing the criterion function over the set of all possible designs. Whether we minimize or maximize depends on the particular criterion function being used. The design that optimizes the criterion function is termed the optimal design. Design criterion functions are tailored for different purposes such as testing, estimation or prediction. Here we consider Bayesian optimal designs for both estimation and testing when a fixed number of observations are to be collected.

Optimal design has been used for many types of models and experiments such as computer experiments (Xu (1999)), blocked experiments (Goos et al. (2005)), cross over models (Matthews (1987)) and regression models (O'Brien and Funk (2003)). Optimal design originated in the regression setting and much has been written about it in this context. Therefore, in this review, most of our references are to regression modelling.

We begin with a brief summary of classical frequentist optimal design, in which Kiefer played a major role. Elfving (1952) introduced optimal design for regression. Shortly after, the alphabetic nomenclature system, currently used in optimal design, was introduced by Kiefer (1958). Many of the criterion functions described by the alphabetic nomenclature are tailored for parameter estimation. The basic idea is to minimize the variability of the estimators for the parameters in the model. Since the asymptotic covariance matrix for the MLEs of the parameters in a regression model is proportional to the inverse of the Fisher information matrix, many of the frequentist design criterion functions for estimation are based on the Fisher information matrix of the model parameters. The most studied design criterion function is D-optimality which maximizes the determinant of the Fisher information matrix. For linear regression, maximizing this determinant is equivalent to minimizing the determinant of the dispersion matrix of the coefficient estimates. Intuitively, D-optimality entails minimization of the volume of the confidence ellipsoid for the model parameters of a given level.

Another design criterion function based on the Fisher information matrix is A-optimality. A-optimality minimizes the trace of the inverse of the Fisher information matrix. That is, the average variance of the parameter estimates is minimized. An example of a frequentist design criterion function for estimation in regression that is not based on the Fisher information matrix is I_L-optimality. Instead, I_L-optimality is based on predicted variance (see Dette and O'Brien (1999)). A design criterion function for model discrimination is T-optimality, see Atkinson and Fedorov (1975).

T-optimality is based on a non-centrality parameter.

Kiefer and Wolfowitz (1959) concentrate on D-optimality in regression. They introduce a design measure, which simplifies the problem. Instead of optimizing the criterion function directly as a function of the design points, one optimizes the criterion function as a function of the design measure. In regression problems, this design measure has the following form: if a total of N measurements is taken and n_i is the integral number of measurements at location x_i, then we assign the weight n_i/N to x_i. With the n_i all integer-valued we have an exact design. For optimization purposes, when we do not insist that the n_i be integer-valued, we may use the approximate continuous theory proposed by Kiefer (1974). Kiefer (1974) also presents the general equivalence theorem. The general equivalence theorem uses convexity results and directional derivatives to describe several equivalent ways in which one can identify the optimal design for concave criterion functions. Often it is much simpler to use the approximate continuous theory and equivalence theorem than to find the optimal design directly, since the latter involves discrete optimization. The first work on equivalence was presented by Kiefer and Wolfowitz (1960). Kiefer and Wolfowitz (1960) demonstrated the equivalence of D-optimality and G-optimality, also known as the minimax criterion. Whittle (1973), considered the equivalence theorem for the non-linear case. His results are included in Kiefer (1974).

Many books have been written about frequentist optimal design in the context of regression. The monograph by Slivery (1980) provides a concise introduction when the underlying model is known. For more recent and in-depth coverage one may consult Pukelsheim (1993), which provides a theoretical view or Atkinson and Donev (1992), which gives a more applied view. Other books on optimal design include Pazman (1986) and Fedorov (1972). A-optimality was extended to the linear optimality criterion by Fedorov (1972) and a corresponding equivalence theorem was found.

The subject of Bayesian optimal design has its origins in designs that are optimal for prediction. See the papers by Lindley (1956) and Lindley (1968), who initially suggested the use of Bayesian optimal design for prediction, and Brooks (1972), Brooks (1974) and Brooks (1976), who examined optimal design for prediction in a regression setting.

The monograph by Pilz (1991) is concerned entirely with Bayesian optimal design for linear regression. Included in this monograph is the aforementioned author's work from the 1980's. Chaloner (1984), El-Krunz and Studden (1991) and Dette (1993) all revisit Elving's original 1952 regression paper in a Bayesian setting. An example of the use of equivalence theory in the Bayesian setting is given in Chaloner and Larntz (1989) who use the equivalence theorem of Whittle (1973) to find the Bayesian optimal design for logistic regression experiments.

Lindley (1972) [pp. 19-20] suggested a Bayesian decision-theoretic framework for Bayesian optimal design. Bernardo and Smith (1994), Berger (1985) and Robert (2001) are all good references for decision theory in a Bayesian statistical context. The review article by Chaloner and Verdinelli (1995) expands on Lindley's idea and places all the work on Bayesian optimal design up until 1995 in the decision-theoretic setting. At the core of decision theory is the utility function or, equivalently, the loss function. These functions are described in Chapter 2 of this thesis and also in Rukhin (1988). As pointed out in Lindley (1972), the two most popular loss functions used for estimation are the Shannon information, introduced by Shannon (1948), and squared error loss. The most popular loss function for testing is the 0-1 loss. Another option for model discrimination is the discrete Spezzaferri criterion function introduced by Spezzaferri (1988). Chaloner and Verdinelli (1995) forge links between the Bayesian design criterion functions and the frequentist alphabetic design criterion functions, where possible. In the linear regression setting they link the Bayes criterion based on the Shannon information loss with D-optimality. They also link the continuous criterion function of Spezzaferri (1988) for estimation to D-optimality. In the same way, A-optimality is linked to a Bayes criterion function based on squared error loss.

Another review article on Bayesian optimal design is contained in the technical report by Clyde (2001). Pukelsheim (1993) and Atkinson and Donev (1992) also both devote a chapter to Bayesian optimal design.

To conclude this section, we point out that in certain situations it is appropriate to combine design criterion functions. One such situation arises when there is model uncertainty; see, for example, Lauter (1974) where a linear combination of design criterion functions is taken for a fixed number of different models. In this case the same type of design criterion function is used for each model. Lauter's work was extended to the Bayesian setting by Zhou et al. (2003). Design criterion functions might also be combined when we are certain about the model but have more than one objective in mind. In this situation it is often appropriate to take a linear combination of two or more design criterion functions for the same model. As an example, if one is interested in both estimation and testing one can take a linear combination of a design criterion function for estimation and a design criterion function for testing and weight them accordingly. Such situations are discussed in Cook and Wong (1994) in the frequentist setting and extended to the Bayesian setting by Clyde and Chaloner (1996).

1.3 Optimal Design in this Thesis

In this thesis we use design criterion functions developed in the Bayesian decision-theoretic framework described by Chaloner and Verdinelli (1995) and Clyde (2001). Specifically, we use the Bayes risk based on squared error loss for estimation and the Bayes risk based on generalized 0-1 loss and Spezzaferri criterion function for model discrimination. These criterion functions are developed in detail in Chapter 2.

Our goal then is to minimize the design criterion functions over the set of allowable designs to find the optimal designs. To do the minimization we introduce a design measure, similar in spirit to the measure introduced by Kiefer for his continuous approximation theory in regression. For any two adjacent design points, our design measure allows us to write a design criterion function in terms of the probability that a change will occur between the points rather than in terms of the distance between those points. The key is that, expressed in terms of the design measure, our criterion functions are all concave functions. The concavity of the criterion functions as a function of the design measure enables us to reduce our original hard optimization problem to a convex optimization problem.

1.3.1 Outline of Thesis

We conclude this chapter by presenting a brief overview of the thesis.

Chapter 2: We present an introduction to decision theory and describe the Bayes risk based on squared error loss, the Bayes risk based on generalized 0-1 loss and the Spezzaferri criterion function for model discrimination. We conclude the chapter by relating the Bayes risk based on the squared error loss to the Spezzaferri criterion.

Chapter 3: We review results from convex optimization, differential geometry, and NEFs needed for this thesis.

Chapter 4: We introduce the Bayesian single-path changepoint model. Next, we present the set of all designs having design points a minimum distance d apart and prove this set forms a simplex. The design measure is introduced.

Chapter 5: We examine the set of allowable design measures. In particular, we focus on how the set of allowable design measures depends on the prior density for the changepoint τ. We conclude by presenting theorems regarding the optimal designs of any design criterion functions which are concave functions of the design measure.

Chapter 6: The Bayesian single-path changepoint model and its posterior distributions are re-expressed in terms of the design measure. The optimal designs for testing for a change, testing for a change in a subinterval, and estimating the before-and-after-change means are all examined. The design criterion functions for these problems are all proved to be concave functions of the design measure. Optimal design results follow from the theorems presented in Chapter 5.

Chapter 7: Examples of the single-path changepoint problem are presented. The first example models observations that are conditionally independent and from any NEF. The second example is the common changepoint multi-path problem which becomes a single-path changepoint problem when the average of the measurements (over paths) at each design point, is taken. Numerical simulations of the Bayes risk based on squared error loss for estimating the before-and-after-change means in the NEF example and the before-and-after-change hierarchical means in the common changepoint multi-path example are given.

Chapter 8: The multiple changepoints multi-path problem is presented. We consider design criterion functions based on squared error loss Bayes risk for estimating the proportion of subjects who do not change and the proportion of subjects who change in a subinterval. We also investigate design criterion functions based on squared error loss Bayes risk for estimating the before-and-after-change hierarchical means. We find that the design criterion functions for these problems are not necessarily concave in the design measure. The non-concavity greatly complicates the problem of finding the optimal design.

Chapter 9: We give some concluding remarks and directions for future work.

Chapter 2

Design Criterion Functions

In this chapter we introduce the three design criterion functions that are used throughout the thesis to find optimal designs. The three criterion functions are the Bayes risk based on generalized 0-1 loss, the Spezzaferri model discrimination criterion function, and the Bayes risk based on squared error loss. In Chapter 6 we use the Bayes risk based on generalized 0-1 loss and the Spezzaferri criterion function for finding optimal designs to test for a change and to test for a change in a subinterval for the single-path model. Likewise, we use a Bayes risk based on squared error loss to find the optimal design when estimation of the before-and-after-change means is the goal. The Bayes risk based on squared error loss is the only design criterion function we consider in Chapter 8. In the multi-path setting of Chapter 8 we write out the design criterion functions for estimating: the proportion of subjects who do not change; the proportion of subjects who change in a subinterval; the before-and-after-change hierarchical means. Since all three criterion functions are motivated from a decision-theoretic point of view, we begin our first section with an introduction to decision theory.

2.1 Decision Theory

Decision theory has been developed to quantify decision making processes when the outcome is uncertain. It plays a major role in the fields of economics, business, statistics, engineering, and game theory. There are both frequentist (classical) and Bayesian approaches to decision theory. As we are concerned with the Bayesian paradigm, we present the Bayesian approach. The books by Berger (1985), Robert (2001) and Bernardo and Smith (1994) cover Bayesian decision theory well. To simplify notation, we do not distinguish between a random variable and the realization of the random variable; lower case letters are used for both, with the interpretation clear from the context.

To introduce the main components of a decision-theoretic problem, let us consider a simple hypothetical example where we wish to estimate a *parameter* θ in the *parameter space* Θ. Our model consists of a density $f(y|\theta)$, which describes how our random observation y is distributed given the unknown parameter θ, and a prior distribution $p(\theta)$, that incorporates our prior beliefs and uncertainties about θ. In our example, the *action space* $\mathcal{A}$ consists of all possible values that we consider to estimate the parameter, and the particular *action* $a \in \mathcal{A}$ is the value we use to estimate θ; that is, the action space $\mathcal{A}$ equals the parameter space Θ. Of course, $\mathcal{A}$ is not always equal to a parameter space. Consider, for instance, a hypothesis testing problem where we either accept or reject a given hypothesis. If we denote accepting by a_0 and rejecting by a_1 then our action space $\mathcal{A}$ would be $\{a_0, a_1\}$.

Statisticians usually base their decision on a *loss function* $L\colon \Theta \times \mathcal{A} \to \mathbb{R}$. In our example $L(\theta, a)$ would represent the loss when action a is taken to estimate θ. Areas such as economics and game theory have an equivalent to the loss function called the *utility function* $U\colon \Theta \times \mathcal{A} \to \mathbb{R}$. So, in the example, instead of quantifying the loss incurred when action a is taken to estimate θ, the value of the utility function

$U(\theta, a)$ describes the gain when action a is taken to estimate θ. The loss function is essentially a negative utility function. As we shall see, interest usually lies in either minimizing the expected loss or in maximizing the expected utility.

Bernardo and Smith (1994) provide a rigorous justification of utility functions. Robert (2001) and Berger (1985) also justify the existence of utility functions. Moreover, Berger (1985) has a fairly lengthy discussion concerning how a utility function can be made into a loss function. Transforming a utility function into a loss function is complicated by the fact that utility functions often have a different domain than $\Theta \times \mathcal{A}$, the domain of their corresponding loss function.

Example 2.1. Squared Error Loss

A loss function which is often used for estimation is the squared error loss

$$L(\theta, a) = (\theta - a)^2.$$

Example 2.2. Generalized 0-1 Loss

A loss function often used for hypothesis testing is the generalized 0-1 loss. Say we have a null hypothesis H_0 and an alternative hypothesis H_1 concerning the value of θ. We partition the Ω space of θ into the events E_0 and E_1 such that the event E_0 corresponds to the truth of H_0 and the event E_1 corresponds to the truth of H_1. The value of a constant K_0 is selected to quantify the loss incurred when H_0 is chosen but the true state is E_1. Likewise, the value of K_1 is selected to quantify the loss incurred when H_1 is chosen but the true state is E_0. Letting I denote the indicator function and taking a_0 to denote that we accept H_0 and a_1 to denote that we accept H_1, we have

$$L(\theta, a_0) = K_0 \, I_{\theta \in E_1},$$

$$L(\theta, a_1) = K_1 \, I_{\theta \in E_0}.$$

A *decision rule* is a mapping $\delta \colon \mathcal{Y} \to \mathcal{A}$. In our example the decision rule is a function on the *sample space* $\mathcal{Y}$ of $f(y|\theta)$ to the parameter space Θ. Once the data y have been collected, $\delta(y)$ takes the value a to estimate θ.

Definition 2.1. The *risk function* of a decision rule $\delta(y)$ is defined as

$$R(\theta, \delta) = E[L(\theta, \delta)] = \int_{\mathcal{Y}} L(\theta, \delta(y)) f(y|\theta) dy.$$

In general, both Bayesians and frequentists want decision rules to provide small values for the risk $R(\theta, \delta)$, which is the average loss, over all the anticipated data, for a given θ. However, θ is unknown and hence it is difficult to provide a decision rule δ such that the risk is small. In the Bayesian setting, we have an advantage because we can use the prior distribution $p(\theta)$ to compute an average risk (this time over θ). This doubly-averaged risk is known as a *Bayes risk.*

Definition 2.2. The *Bayes risk* of a decision rule δ, with respect to a prior distribution $p(\theta)$, is defined as

$$r(p, \delta) = E[R(\theta, \delta)] = \int_{\Theta} R(\theta, \delta) p(\theta) d\theta.$$

Obviously, we would like to obtain the decision rule that minimizes the Bayes risk. Such a decision rule is called the *Bayes rule.*

Definition 2.3. The *Bayes rule* δ^p is the decision rule that minimizes the Bayes risk $r(p, \delta)$.

Example 2.3. Bayes Rule for Squared Error Loss
The Bayes rule for the squared error loss Bayes risk in Example 2.1 is the posterior expectation $E(\theta|y)$.

Example 2.4. Bayes Rule for Generalized 0-1 loss
For the hypothesis testing situation, letting $\mathcal{A} = \{a_0, a_1\}$, *when* $\delta(y) = a_0$ *we accept the null hypothesis* H_0 *and when* $\delta(y) = a_1$ *we reject. To form the decision rule, we partition* $\mathcal{Y}$ *into* R_0 *and* R_1 *such that* $y \in R_0$ *implies that* $\delta(y) = a_0$ *and* $y \in R_1$

implies that $\delta(y) = a_1$. A quick calculation shows that the Bayes rule is

$$\begin{aligned} y \in R_0 \quad \text{if} \quad & K_0 \int_{\theta \in E_1} f(y|\theta)p(\theta)d\theta < K_1 \int_{\theta \in E_0} f(y|\theta)p(\theta)d\theta \\ y \in R_1 \quad \text{if} \quad & K_1 \int_{\theta \in E_0} f(y|\theta)p(\theta)d\theta < K_0 \int_{\theta \in E_1} f(y|\theta)p(\theta)d\theta. \end{aligned}$$

2.2 Bayes Risks as Design Criterion Functions

The Bayes risk design criterion functions used in this thesis are simply the Bayes risks based on the appropriate squared error and generalized 0-1 losses. In this form, the Bayes risk provides the decision rule that minimizes its value. For optimal design purposes, we consider each Bayes risk as a function of the design and find the design providing the lowest value of the risk.

In a Bayesian optimal design problem for estimation, we must average over both the anticipated data and the values of the parameter. The resulting risk must then be minimized as a function of the various designs. Suppressing the dependence on the design, the Bayes risk based on the squared error loss is

$$\int \int \left(\theta - E(\theta|y)\right)^2 f(\theta|y) f(y) d\theta dy. \tag{2.1}$$

Expression (2.1) is a design criterion function usually used for estimation, though as we will see shortly, it is related to the Spezzaferri criterion for model discrimination. It is easily seen that expression (2.1) simplifies to

$$\int Var(\theta|y) f(y) dy. \tag{2.2}$$

In Section 1.2.2 we mentioned that the Bayes risk based on the squared error loss is commonly referred to as the Bayesian A-optimality criterion. The reason is that in regression there is a strong similarity between the frequentist A-optimality design criterion function (the trace of the inverse of the Fisher information matrix) and the Bayes risk based on squared error loss. For our changepoint model, the expressions for the trace of the inverse Fisher information matrix and for the Bayes risk based on squared error loss are not, in fact, similar.

The Bayes risk based on generalized 0-1 loss is

$$K_0 \int_{R_0} \int_{\theta \in E_1} f(y|\theta)p(\theta)d\theta dy + K_1 \int_{R_1} \int_{\theta \in E_0} f(y|\theta)p(\theta)d\theta dy. \tag{2.3}$$

Expression (2.3) is a design criterion function for hypothesis testing discussed by Felsenstein (1992) and used by Blackmore and Williams (2005).

2.3 Scoring Rule Utilities

The Spezzaferri criterion (Spezzaferri (1988)) is based on a special type of utility function called a *scoring rule.* Scoring rules are discussed in detail by Bernardo and Smith (1994) and more recently by Gneiting and Raftery (2004). This short introduction to scoring rules is largely based on the aforementioned reference by Bernardo and Smith (1994). Taking a Bayesian viewpoint, suppose we represent our beliefs about the truth of a set of hypotheses $\{H_j, j \in J\}$ by a distribution $\{q_j, j \in J\}$. To do so, we partition the model space into the events $\{E_j, j \in J\}$, where event E_j is equivalent to the hypothesis H_j being true. We represent all possible distributions over $\{E_j, j \in J\}$ as

$$\mathcal{Q} = \{q \equiv \{q_j, j \in J\}; q_j \geq 0, \sum_{j \in J} q_j = 1\}.$$

Scoring rules quantify the gain when using a distribution q to represent the belief about the truth of statements $\{E_j, j \in J\}$.

Definition 2.4. A *scoring rule* U for probability distributions $q = \{q_j, j \in J\}$ defined over a partition $\{E_j, j \in J\}$ is a mapping which assigns a real number $U(q, E_j)$ to each pair (q, E_j).

Usually, the goal is to find a distribution $q \in \mathcal{Q}$ which maximizes the expected utility,

$$\sum_{j \in J} U(q, E_j) P(E_j | y). \tag{2.4}$$

The true current belief of a Bayesian is the posterior distribution $\{P(E_j|y), j \in J\}$; therefore we would like a scoring rule whose expectation (2.4) is maximized when q equals the posterior distribution for any given y. Such utilities are called *proper scoring rules.*

Definition 2.5. A scoring rule U is *proper* if, for each probability distribution $p = \{p_j, j \in J\}$ defined over a partition $\{E_j, j \in J\}$,

$$\sup_{q \in \mathcal{Q}} \left\{ \sum_{j \in J} U(q, E_j) p_j \right\} = \sum_{j \in J} U(p, E_j) p_j$$

and the supremum is attained if and only if $q = p$.

The Spezzaferri criterion function is based on a particular proper scoring rule introduced by Brier (1950) in the context of weather forecasting, and again later by DeFinetti (1962). It is called the *quadratic scoring rule.*

Definition 2.6. A *quadratic scoring rule* for probability distributions $q = \{q_j, j \in J\}$ defined over a partition $\{E_j, j \in j\}$ is any function of the form

$$U(q, E_j) = A\big(2q_j - \sum_{i \in J} q_i^2\big) + B_j, \;\; A > 0.$$

It is easily shown that the quadratic scoring rule is proper. It is also easily shown that the quadratic score function can be rewritten as

$$U(q, E_j) = A\big(1 - \sum_{i \in J} (q_i - I_{E_j})^2\big) + B_j, \;\; A > 0. \tag{2.5}$$

2.4 The Spezzaferri Criterion

The Spezzaferri criterion (Spezzaferri (1988)) was introduced for both estimation and model discrimination. To construct a criterion function for model estimation, Spezzaferri defined a continuous analogue to the discrete quadratic scoring rule. In this thesis we use the Spezzaferri criterion for model discrimination and therefore rederive it using the original discrete form of the quadratic scoring rule. The Spezzaferri criterion function has been advocated as a Bayesian criterion function for model discrimination by both Chaloner and Verdinelli (1995) and Clyde (2001).

This criterion posits that the usefulness of an experiment e is measured by the expected relative increase of utility after the experiment is performed. Letting U denote the quadratic scoring rule, in the hypothesis testing situation $J = \{0, 1\}$ we have

$$U(e) = \frac{\sum_{j=0}^{1} \big[U(P(E_j|y), E_j) - U(P(E_j), E_j)\big] P(E_j|y)}{\sum_{j=0}^{1} U(P(E_j), E_j) P(E_j)}. \tag{2.6}$$

Obviously the denominator does not depend on the data and hence has no effect on the design. Consequently, we concern ourselves with the numerator. By substituting expression (2.5) into the numerator of (2.6), we obtain

$$\begin{aligned}
&A\big[P(E_0)^2 - P(E_0|y)^2 + P(E_1)^2 - P(E_1|y)^2 \\
&+ 2(P(E_0|y) - P(E_0))\big] P(E_0|Y) \\
&+ A\big[P(E_0)^2 - P(E_0|y)^2 + P(E_1)^2 - P(E_1|y)^2 \\
&+ 2(P(E_1|y) - P(E_1))\big] P(E_1|Y).
\end{aligned}$$

Upon simplification, we find for the numerator of $U(e)$

$$\begin{aligned}
&A\big[(P(E_0)^2 + P(E_1)^2 + P(E_0|y)^2 + P(E_1|y)^2 \\
&- 2(P(E_0)P(E_0|y) + P(E_1)P(E_1|y))\big].
\end{aligned}$$

The Spezzaferri criterion function is formed by averaging $\int U(e) f(y) dy$ over the anticipated data, $f(y)$. The term $2\int \big(P(E_0)P(E_0|y) + P(E_1)P(E_1|y)\big) f(y) dy$ is constant and equal to $2\big(P(E_0)^2 + P(E_1)^2\big)$. Consequently, maximizing the Spezzaferri criterion function $\int U(e)f(y)dy$ is equivalent to maximizing

$$\int \big(P(E_0|y)^2 + P(E_1|y)^2\big) f(y)dy. \tag{2.7}$$

Using the relation

$$\big(P(E_0|y) + P(E_1|y)\big)^2 - 2P(E_0|y)P(E_1|y) = P(E_0|y)^2 + P(E_1|y)^2,$$

it is easily seen that maximizing (2.7) is equivalent to minimizing

$$\int P(E_0|y)P(E_1|y)f(y)dy \tag{2.8}$$

which equals

$$\int P(E_0|y)^2 P(E_1|y)f(y)dy + \int P(E_1|y)^2 P(E_0|y)f(y)dy. \tag{2.9}$$

All the above observations were made by Spezzaferri (1988).

2.5 Comments on the Spezzaferri Criterion

Our final section in this chapter relates the Spezzaferi criterion function to the Bayes risk based on squared error loss and to the posterior expectation (2.4) of the quadratic score utility. To our knowledge, these observations have not been made before. We believe such comparisons, although simple, are important, as it is crucial to understand all possible interpretations of a particular design criterion function. As we will see at the end of this section, our alternative interpretations of the Spezzaferri criterion function suggest how a hypothesis test should be conducted when the Spezzaferri criterion function is used.

Theorem 2.7. *The testing design criterion function formed by finding the design which maximizes the expected quadratic score utility $\sum_{j=0}^{1} U(p, E_j)p_j$, with $B_0 = B_1$ in equation (2.5) and $p = \{P(E_j|y), j \in \{0,1\}\}$, is equivalent to minimizing (2.9) or, equivalently, (2.8).*

Proof. Letting U denote the quadratic scoring rule, we first calculate

$$U(e) = \sum_{j=0}^{1} U(P(E_j|y), E_j)P(E_j|y)$$

and then we average over the data.

Using expression (2.5) we find that,

$$\begin{aligned} U(e) =& A + B_0 P(E_0|y) + B_1 P(E_1|y) \\ & - 2AP(E_1|y)^2 P(E_0|y) - 2AP(E_0|y)^2 P(E_1|y). \end{aligned}$$

Taking $B_0 = B_1 = B$ we have

$$U(e) = A + B - 2AP(E_1|y)^2 P(E_0|y) - 2AP(E_0|y)^2 P(E_1|y).$$

Once averaged over $f(y)$, maximizing $\int U(e)f(y)dy$ is equivalent to minimizing $\int P(E_0|y)^2 P(E_1|y)f(y)dy + \int P(E_1|y)^2 P(E_0|y)f(y)dy$, that is expression (2.9). □

Immediately following Theorem 2.7, we see that using the Spezzaferri criterion function to discriminate between two models is equivalent to maximizing the average posterior expectation of the quadratic scoring rule with $B_0 = B_1$.

Now consider the Bayes risk based on squared error loss. As we mentioned, this criterion function is usually used for estimation. However, we can think of the model discrimination problem as one of estimation, of I_{E_j}. Since the Bayes rule is the posterior expectation of the parameter of interest, the Bayes estimator for I_{E_j} is the posterior probability $P(E_j|y)$.

Theorem 2.8. *The Bayes risk based on squared error loss design criterion function for estimating I_{E_0} is equivalent to the Bayes risk based on squared error loss design criterion function for estimating I_{E_1}. Furthermore, both these criterion functions are equivalent to finding the design which minimizes (2.9) or, equivalently, (2.8).*

Proof. Obviously, proving the second statement of the theorem is enough to certify the truth of the first. We prove it for I_{E_0}. By symmetry, the proof for I_{E_1} follows. The Bayes risk based on squared error loss for I_{E_0} is

$$\int \left(1 - P(E_0|y)\right)^2 P(E_0|y) f(y) dy + \int \left(0 - P(E_0|y)\right)^2 P(E_1|y) f(y) dy$$

which immediately reduces to $\int P(E_0|y)^2 P(E_1|y) f(y) dy + \int P(E_1|y)^2 P(E_0|y) f(y) dy$, that is, expression (2.9). □

Now that we have shown the equivalence of the three criterion functions (Spezzaferri, squared error loss Bayes risk, and expected quadratic score), we can discuss how a hypothesis test should be conducted when the Spezzaferri criterion function has been used. Theorem 2.8 makes this suggestion obvious. Since the Spezzaferri criterion function is one which allows us to find the design to "best" estimate I_{E_0} and I_{E_1} by the posterior densities $P(E_0|y)$ and $P(E_1|y)$, our decision should be based on one or both of the values $P(E_0|y)$ and $P(E_1|y)$. For instance, as suggested by Casella and Berger (2002) [p. 397], the decision could be made to accept H_0 if

$$P(E_0|y) > P(E_1|y)$$

and reject H_1 otherwise. Notice this is the Bayes rule for the Bayes risk based on 0-1 loss. Casella and Berger (2002) [p. 397] propose that one can guard against falsely rejecting H_0 by deciding to reject H_0 only if $P(E_1|y)$ is greater than a large number such as 0.99.

Chapter 3

Preliminaries

This third chapter provides the necessary results from convex optimization, differential geometry, and natural exponential families (NEFs) to read this thesis. Section 3.1 covers convex optimization, Section 3.2 covers differential geometry and Section 3.3 covers NEFs and their Diaconis-Ylvisaker-conjugate or DY-conjugate prior distributions. The results from convex optimization and differential geometry are used throughout the thesis. The properties of NEFs are used in Chapter 7, in an example of a single-path changepoint model. Where possible, we indicate how the results are used in the sequel.

3.1 Convex Optimization

We begin with some important results from convex optimization, which can be found in any book on convex optimization or convex analysis; see, for example, the books by Ben-Tal and Nemirovskii (2001) and Rockafellar (1970). The following is largely based on Chapters 1, 2, 4 and 5 of the book by Ben-Tal and Nemirovski.

To simplify notation, we do not use special characters to differentiate vectors and scalars. The dimensionality of any particular quantity should be evident from the context. We denote the usual standard basis of $\mathbb{R}^n$ by $\{e_1, \ldots, e_n\}$. The zero vector is denoted by e_0.

An *affine combination* of the set of vectors $\{v_0, \ldots, v_k\} \subset \mathbb{R}^n$ is a linear combination $\sum_{i=0}^k \lambda_i v_i$ such that $\sum_{i=0}^k \lambda_i = 1$.

A set of vectors $\{v_0, \ldots, v_k\} \subset \mathbb{R}^n$ is *affinely independent* if there does not exist a non-trivial linear combination which equals zero and that has coefficients which sum to zero.

Definition 3.1. A collection $\{v_0, \ldots, v_k\} \subset \mathbb{R}^n$ of vectors is *affinely independent* if

$$\sum_{i=0}^{k} \lambda_i v_i = 0, \sum_{i=0}^{k} \lambda_i = 0 \Longrightarrow \lambda_0 = \cdots = \lambda_k = 0.$$

Example 3.1. Affine Independence

The vectors e_0, e_1 and e_2 in $\mathbb{R}^2$ are affinely independent. More generally, the vectors $e_0, e_1, \ldots, e_n$ are affinely independent in $\mathbb{R}^n$. Notice that affine independence does not imply linear independence.

An important consequence of Definition 3.1 is the following lemma.

Lemma 3.1. *Let $\{v_0, \ldots, v_n\} \subset \mathbb{R}^n$ be an affinely independent set of vectors. Then the coefficients λ_i in an affine combination $v = \sum_{i=0}^k \lambda_i v_i$, $\sum_{i=0}^k \lambda_i = 1$, of $\{v_0, \ldots, v_k\}$ are uniquely determined by v.*

The values of the coefficients λ_i in an affine combination are termed the *barycentric coordinates* and, as such, we refer to the vector $(\lambda_0, \ldots, \lambda_n)$ as the *barycentric vector.*

Example 3.2. Unique Barycentric Coordinates

Consider the affine combination $\lambda_0 e_0 + \lambda_1 e_1 + \lambda_2 e_2$ in $\mathbb{R}^2$. Since e_0, e_1 and e_2 are affinely independent there is no other affine combination of e_0, e_1 and e_2 whose barycentric coordinates are $(\lambda_0, \lambda_1, \lambda_2)$. Note that $\lambda_0 = 1 - \lambda_1 - \lambda_2$. More generally, the affine combination $\lambda_0 e_0 + \cdots + \lambda_n e_n$ has the barycentric coordinates $(\lambda_0, \ldots, \lambda_n)$

where $\lambda_0 = 1 - \lambda_1 - \cdots - \lambda_n$, and there is no other affine combination of $e_0, \ldots, e_n$ whose barycentric coordinates are $(\lambda_0, \ldots, \lambda_n)$.

Another consequence of Definition 3.1 is Lemma 3.2.

Lemma 3.2. *The set of vectors $\{v_0, \ldots, v_k\}$ are affinely independent if and only if the k vectors $(v_1 - v_0), \ldots, (v_k - v_0)$ are linearly independent.*

Example 3.3. Affine Independence and Linear Independence
As we saw in Example 3.1 the vectors e_0, e_1 and e_2 in $\mathbb{R}^2$ are affinely independent. By Lemma 3.2 we note that the vectors $e_1 - e_0$ and $e_2 - e_0$ are linearly independent. Similarly, the vectors $e_0, e_1, \ldots, e_n$ in $\mathbb{R}^n$ are affinely independent and the vectors $e_1 - e_0, \ldots, e_n - e_0$ are linearly independent.

Throughout the thesis, to differentiate between a vector component z_a of a vector z and a vector z with a subscript a, we denote a vector z with a subscript a as $z_{(a)}$.

Definition 3.2. A set $B \subseteq \mathbb{R}^n$ is *convex* if

$$z_{(a)},\, z_{(b)} \in B,\ t_a,\, t_b \geq 0,\ \ t_a + t_b = 1 \Longrightarrow t_a z_{(a)} + t_b z_{(b)} \in B.$$

In other words, B contains any line segment joining two elements of B.

Definition 3.3. A *convex combination* of vectors $\{v_0, \ldots, v_k\} \subset \mathbb{R}^n$ is

$$v = \sum_{i=0}^{k} \lambda_i v_i \text{ s.t. } \lambda_i \geq 0\ \ \forall\, i,\ \sum_{i=0}^{k} \lambda_i = 1.$$

Given $B \subseteq \mathbb{R}^n$, not necessarily convex, then the *convex hull* of B, denoted $\mathrm{Conv}(B)$ is the smallest convex set containing B. Equivalently, $\mathrm{Conv}(B)$ is the intersection of all convex sets containing B. It also happens that for any non-empty set B in $\mathbb{R}^n$ that

$$\mathrm{Conv}(B) = \{\text{the set of all convex combinations of vectors from } B\}$$

The particular type of convex set that interests us is a *simplex*. An n-dimensional simplex is the convex hull of $n+1$ affinely independent points $\{v_0, \ldots, v_n\} \subset \mathbb{R}^n$. The points $\{v_0, \ldots, v_n\}$ are called the *vertices* of the simplex. In three dimensions or less, a simplex is easily visualized: a *one*-dimensional simplex is a line, a *two*-dimensional simplex is a triangle, and a *three*-dimensional simplex is a tetrahedron.

Definition 3.4. A *simplex* with affinely independent vertices $\{v_0, \ldots, v_n\} \subset \mathbb{R}^n$ is a convex set B such that

$$B = \mathrm{Conv}(v_0, \ldots, v_n) = \{\sum_{i=0}^{n} \lambda_i v_i \mid \lambda_i \geq 0 \ \ \forall\ i,\ \sum_{i=0}^{n} \lambda_i = 1\}.$$

Since the barycentric coordinates λ of each point in an n-dimensional simplex are all positive and sum to one, they can represent a probability measure assigning mass to a discrete random variable with $n+1$ support points. In Chapters 6, 7, and 8, we re-express our design criterion functions as functions of such a probability measure and find the probability measure which minimizes the design criterion functions. The design criterion functions will then be scalar functions of the barycentric coordinates of *any* n-dimensional simplex. In this thesis we use the simplex S^n which has the vertices $\{e_0, \ldots, e_n\}$. Defined in terms of the Euclidean coordinates, $z = (z_1, \ldots, z_n)$, S^n is the region $z_1 + \cdots + z_n \leq 1$ and $z_i \geq 0$ for all $i = 1, \ldots, n$.

Example 3.4. Convex Hull

The convex hull of the affinely independent points e_0, e_1 *and* e_2 *in* $\mathbb{R}^2$ *is the simplex* S^2 *shown in Figure 3.1. In Euclidean coordinates* S^2 *is defined by* $z = (z_1, z_2)$ *such that* $z_1 + z_2 \leq 1$, $z_1 \geq 0$ *and* $z_2 \geq 0$. *The barycentric coordinates are* $\lambda_0 = 1 - z_1 - z_2$, $\lambda_1 = z_1$, *and* $\lambda_2 = z_2$.

The Euclidean vector $z \in S^n$ is equal to the affine combination $\sum_{i=0}^{n} \lambda_i e_i$ where $\lambda_0 = 1 - z_1 - \cdots - z_n$ and $\lambda_i = z_i$ for all $i = 1, \ldots, n$. By Lemma 3.1, this relationship between z and $\lambda = (\lambda_0, \ldots, \lambda_n)$ is unique. Due to this one-to-one relation, any scalar

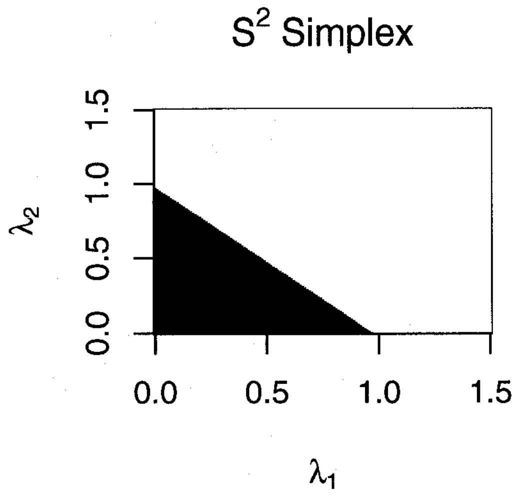


Figure 3.1: The S^2 simplex with vertices e_0, e_1 and e_2.

function over S^n can either be expressed as a function of the Euclidean coordinates z or the barycentric coordinates λ. Considering h and g which both take the same values over S^n, but where h is a function of barycentric coordinates λ and g is a function of Euclidean coordinates z, we see that

$$h(\lambda) = h(\lambda_0, \dots, \lambda_n) = h(1 - z_1 - \dots - z_n, z_1, \dots, z_n) = g(z_1, \dots, z_n) = g(z). \quad (3.1)$$

In Chapter 6 we prove that our design criterion functions are concave when expressed as scalar functions of the barycentric coordinates of S^n. The following definition of concavity is expressed in terms of the Euclidean coordinates.

Definition 3.5. A function $g\colon S^n \to \mathbb{R}$ expressed in terms of the Euclidean coordinates, is *concave* over S^n if for any $z_{(a)}$, $z_{(b)} \in S^n$ and $\forall\, t_a, t_b \geq 0$ such that $t_a + t_b = 1$, we have

$$g(t_a z_{(a)} + t_b z_{(b)}) \geq t_a g(z_{(a)}) + t_b g(z_{(b)}).$$

If a function g is concave then the function $-g$ is *convex*.

Since our design criterion functions are expressed in terms of the barycentric coordinates of S^n, in Lemma 3.3 we re-write Definition 3.5.

Lemma 3.3. *Consider a function $h\colon S^n \to \mathbb{R}$, expressed in terms of the barycentirc coordinates of S^n. If for all $\lambda_{(a)}, \lambda_{(b)} \in S^n$ and $t_a, t_b \geq 0$ with $t_a + t_b = 1$, we have*

$$h(t_a\lambda_{(a)} + t_b\lambda_{(b)}) \geq t_a h(\lambda_{(a)}) + t_b h(\lambda_{(b)}),$$

then h is concave over S^n.

Proof. By equation (3.1) the function $h(\lambda)$ can be re-expressed as a function $g(z)$. Furthermore, from the one-to-one relation in Lemma (3.1) we have that for any $\lambda_{(a)}$ and $\lambda_{(b)}$ there are corresponding $z_{(a)}$ and $z_{(b)}$. It is easily seen that $h(t_a\lambda_{(a)} + t_b\lambda_{(b)}) = g(t_a z_{(a)} + t_b z_{(b)})$, that $t_a h(\lambda_{(a)}) = t_a g(z_{(a)})$, and that $t_b h(\lambda_{(b)}) = t_b g(z_{(b)})$.

Therefore, Lemma 3.3 implies Definition 3.5 holds over the simplex $S^n \subset \mathbb{R}^n$ and vice versa. □

An immediate consequence of Lemma 3.3 is that if h is concave over S^n, then h is also concave over any m-dimensional subspace of S^n formed by setting the components λ_0 through to λ_{n-m} equal to zero.

In Chapter 6, our design criterion functions are in integral form over the data $y = (y_1, \ldots, y_n)$. Hence the following theorem is needed.

Theorem 3.6. *If $h(\lambda; y)$ is concave over S^n (in the sense of Lemma 3.3) for all points y in Ω, then $\int_\Omega h(\lambda; y)dy$ is also concave in S^n (in the sense of Lemma 3.3).*

Proof. Since $h(\lambda; y)$ is concave we have

$$h(t_a\lambda_{(a)} + t_b\lambda_{(b)}; y) \geq t_a h(\lambda_{(a)}; y) + t_b h(\lambda_{(b)}; y)$$

where $t_a + t_b = 1$. It follows immediately that

$$\int_\Omega h(t_a\lambda_{(a)} + t_b\lambda_{(b)}; y)dy \geq t_a \int_\Omega h(\lambda_{(a)}; y)dy + t_b \int_\Omega h(\lambda_{(b)}; y)dy$$

where $t_a + t_b = 1$. Hence $\int_\Omega h(\lambda; y)dy$ is concave. □

Jensen's Inequality, a well-known result in convex optimization, is used in the proof of Theorem 3.8.

Theorem 3.7. Jensen's Inequality

For any concave function $g\colon \mathbb{R}^n \to \mathbb{R}$, *the following is true. For* $\lambda_i > 0$ *for* $i = 1, \ldots, k$ *and* $\sum_{i=0}^{k} \lambda_i = 1$ *we have*

$$g\left(\sum_{i=0}^{k} \lambda_i v_i\right) \geq \sum_{i=0}^{k} \lambda_i g(v_i).$$

The proof of Jensen's Inequality uses strong induction and the definition of concavity.

Theorem 3.8. *Let* C^n *be any* n*-dimensional simplex with vertices* $\{v_0, \ldots, v_n\}$. *A concave function* $g\colon C^n \to \mathbb{R}$ *takes its minimum value at one of the vertices of* C^n.

Proof. Consider any point $z \in C^n$. Then, by the definition of the simplex C^n, this point can be expressed as a convex combination of the set of vertices $\{v_0, \ldots, v_n\}$. Hence by the concavity of g, and using Jensen's Inequality,

$$\begin{aligned} g(z) &= g(\lambda_0 v_0 + \cdots + \lambda_n v_n) \\ &\geq \lambda_0 g(v_0) + \cdots + \lambda_n g(v_n) \\ &\geq \lambda_0 \min\{g(v_0), \ldots, g(v_n)\} + \cdots + \lambda_n \min\{g(v_0), \ldots, g(v_n)\} \\ &= \min\{g(v_0), \ldots, g(v_n)\}. \end{aligned}$$

□

An obvious extension is stated as a corollary below and plays a crucial role when minimizing the design criterion functions.

Corollary 3.1. *A concave function minimized over a subset of a simplex which contains the vertices of the simplex is minimized at one of the vertices of the simplex.*

Lastly we state and prove Theorem 3.9, which is needed in Section 5.6.2 when considering the optimal design for testing for a change in a subinterval. Let H be the subset of S^n which is the product of two simplices H_1 and H_2. In particular we have the barycentric coordinates $\lambda \in S^n$, where the coordinate λ_q and the sums $\sum_{j=0}^{q-1} \lambda_j$, and $\sum_{j=q+1}^{n} \lambda_j$ are fixed. First of all, we restrict ourselves to the $(n-1)$-dimensional hyperplane where λ_q is fixed. Secondly, the barycentric coordinates $(\lambda_0, \ldots, \lambda_{q-1})$ described by the fixed sum $\sum_{j=0}^{q-1} \lambda_j$ are associated with an $(q-1)$-dimensional simplex we denote H_1, while the barycentric coordinates $(\lambda_{q+1}, \ldots, \lambda_n)$ described by the fixed sum $\sum_{j=q+1}^{n} \lambda_j$ are associated with an $(n-(q+1))$-dimensional simplex we denote H_2. If $\hat{\lambda}_q$ indicates omission of λ_q, the points $(\lambda_0, \ldots, \lambda_{q-1}, \hat{\lambda}_q, \lambda_{q+1}, \ldots, \lambda_n)$ lie in the product space $H_1 \times H_2$.

Theorem 3.9. *Consider a concave function $g(z)$ over the product space $H_1 \times H_2$ where H_1 and H_2 are $(q-1)$-dimensional and $(n-(q+1))$-dimensional simplices in S^n. Let $\{a_0, \ldots, a_{q-1}\}$ be the vertex set of H_1 and $\{b_0, \ldots, b_{n-(q+1)}\}$ be the vertex set of H_2. The minimum of $g(z)$ will occur at a point which is the Cartesian product of a vertex in $\{a_0, \ldots, a_{q-1}\}$ with a vertex in $\{b_0, \ldots, b_{n-(q+1)}\}$.*

Proof. Let z_1 be an arbitrary point in H_1 and z_2 be an arbitrary point in H_2. Then for any fixed point w_1 in H_1 the function $g(w_1, z_2)$ is concave over H_2 and for any fixed point w_2 in H_2 the function $g(z_1, w_2)$ is concave over H_1. By Theorem 3.8, for *any* fixed w_1, $g(w_1, z_2)$ will be minimized at one of $\{b_0, \ldots, b_{n-(q+1)}\}$. Similarly, for *any* fixed w_2, $g(z_1, w_2)$ will be minimized at one of $\{a_0, \ldots, a_{q-1}\}$. It follows that the minimum of $g(z)$, with domain $H_1 \times H_2$, is at a point z^* which is a Cartesian product of a point in $\{a_0, \ldots, a_{q-1}\}$ with a point in $\{b_0, \ldots, b_{n-(q+1)}\}$. □

A corollary to Theorem 3.9 is stated below.

Corollary 3.2. *Consider a product space of a $(q-1)$-dimensional volume with a $(n-(q+1))$-dimensional volume. Suppose that the first volume is a subset of a* *simplex H_1 and contains the vertices $\{a_0, \ldots, a_{q-1}\}$ of H_1. Suppose that the second volume is a subset of a different simplex H_2 and contains the vertices $\{b_0, \ldots, b_{n-(q+1)}\}$ of H_2. A concave function over the product space of the two volumes is minimized at a point which is a Cartesian product of a vertex in $\{a_0, \ldots, a_{q-1}\}$ and a vertex in $\{b_0, \ldots, b_{n-(q+1)}\}$.*

3.2 Differential Geometry

Differential geometry is an area of mathematics in which techniques have been developed to allow us to move along a surface and study its shape and other properties. In Chapter 6, for the single-path problem, we have concave criterion functions. By Corollary 3.1, if the volume over which we are minimizing is a subset of a simplex containing the vertices of this simplex, we can easily minimize the concave criterion functions. Hence in Chapter 5 we study the shape of the subset of S^n over which we must minimize the design criterion functions.

To examine the shape of an n-dimensional surface we consider its $(n-1)$-dimensional boundary. Consequently, our interest is to consider $(n-1)$-dimensional surfaces in $\mathbb{R}^n$. Such surfaces are often called hypersurfaces.

Figure 3.2 depicts such a situation when $n = 2$. Here we have a two-dimensional surface. By considering its one-dimensional boundary, we can ascertain the shape of the two-dimensional surface. It is easy to visualize such a situation for $n = 3$. Although impossible to visualize, the idea remains the same for higher dimensions.

We present some of the results and techniques from differential geometry which allow us to examine the $(n-1)$-dimensional surfaces in $\mathbb{R}^n$. Most elementary textbooks in differential geometry consider at most three-dimensional surfaces. The book by Thorpe (1979) presents a good introduction to differential geometry for n-dimensional surfaces. However, Thorpe (1979) considers only n-dimensional surfaces described as

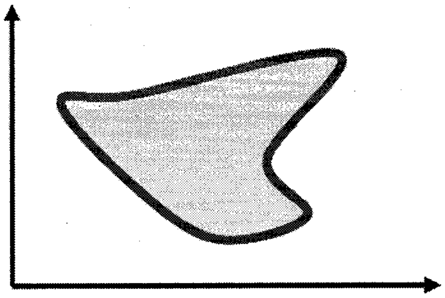


Figure 3.2: A two-dimensional surface with a one-dimensional boundary.

level curves of scalar functions over $\mathbb{R}^n$. The text by O'Neil (1966) considers surfaces described by parametrizations, more in the spirit of this thesis. This section is largely based on Chapters 9, 10, and 12 of Thorpe (1979), appropriately modified to account for the fact that the $(n-1)$-dimensional surfaces we consider are not described as level surfaces of scalar functions, but by their parametrizations. Again, we do not use special notation for vectors, whether or not a quantity is a vector is dictated by the context.

The $(n-1)$-dimensional surfaces in $\mathbb{R}^n$ that we study in Chapter 5 are the image of a set $U \subset \mathbb{R}^{n-1}$ under an injective mapping $Q\colon \mathbb{R}^{n-1} \to \mathbb{R}^n$. Hence, we wish to study the shape of $Q(U) = (Q^1(U), \ldots, Q^n(U))$. We have labelled the components of $Q(U)$ as superscripts to allow ourselves room to take partial derivatives as subscripts later on. Although $Q(U)$ lies in $\mathbb{R}^n$ it only has dimension $n-1$; this is analogous to Figure 3.2, where we observed a one-dimensional boundary lying in $\mathbb{R}^2$.

Our $(n-1)$-dimensional surface $Q(U)$ is parametrized by the coordinates of the points $z = (z_1, \ldots, z_{n-1})$ lying in U. That is,

$$Q(z_1, \ldots, z_{n-1}) = (Q^1(z_1, \ldots, z_{n-1}), \ldots, Q^n(z_1, \ldots, z_{n-1})).$$

It follows that as the coordinates in U change, we move around the surface $Q(U)$. To study $Q(U)$ we make use of 1-dimensional *parametrized curves* lying in $Q(U)$. A curve lying in $Q(U)$ parametrized by t will be a smooth function $\alpha\colon I \to \mathbb{R}^n$, where

I is an open interval in $\mathbb{R}$. That is, $\alpha(t) = (\alpha_1(t), \dots, \alpha_n(t))$. If $\alpha(t_0) = q$, then the derivative $\alpha'(t_0) = (\alpha_1'(t_0), \dots, \alpha_n'(t_0))$ is a vector tangent to $Q(U)$ at q. This derivative is also the velocity vector of the curve $\alpha(t)$ at t_0. Example 3.5 shows such a parametrized curve and its derivative. We denote the tangent space to $Q(U)$ at q by $T_qQ(U)$.

Example 3.5. Parametrized Curves

Let $p \in U$ such that $q = Q(p)$. Then the curve

$$\alpha(t) = Q(p_1, \dots, p_{j-1}, t, p_{j+1}, \dots, p_{n-1})$$

is obviously a curve in $Q(U)$ parametrized by t. The derivative of $\alpha(t)$ evaluated at $t = p_j$ is a vector in the tangent space $T_qQ(U)$. In fact, it is easy to see that $\alpha'(p_j)$ equals the partial derivative $Q_{z_j} = (Q^1_{z_j}, \dots, Q^n_{z_j})$ evaluated at p. Hence, the partial derivative $Q_{z_j}|_p$ is a vector in $T_qQ(U)$.

Obviously there are $n-1$ such tangent vectors in $T_qQ(U)$ corresponding to the $n-1$ partial derivatives $Q_{z_j}|_p$. If the Jacobian of Q is one-to-one, these vectors are linearly independent. The curve in Example 3.5 is the type of curve we use to obtain our results in Chapter 5.

Parametrized curves can be used when computing *directional derivatives.* The definition of a directional derivative for both a scalar function and a mapping defined on $Q(U)$ is given below.

Definition 3.10. A directional derivative of a scalar function $f \colon Q(U) \to \mathbb{R}$, in the direction of the n-dimensional vector v at the point $q \in Q(U)$, is defined as

$$\nabla_v f = \nabla f|_q \cdot v.$$

The directional derivative of a mapping $f \colon Q(U) \to \mathbb{R}^n$, in the direction of the n-dimensional vector v at the point q, is defined as

$$\nabla_v f = (\nabla_v f^1, \dots, \nabla_v f^n).$$

If $\alpha\colon I \to Q(U) \subset \mathbb{R}^n$ is such that $\alpha(t_0) = q$ and $\alpha'(t_0) = v$, the chain rule can be used to show that

$$(f \circ \alpha)'(t_0) = \nabla f(\alpha(t_0)) \cdot \alpha'(t_0) = \nabla f|_q \cdot v. \tag{3.2}$$

Recall, our goal is to investigate the shape of an $(n-1)$-dimensional surface $Q(U)$ in $\mathbb{R}^n$. The general idea is to observe how a normal vector to $Q(U)$ changes as one moves around the surface $Q(U)$. Hence, we first contemplate how to calculate normal vectors to $Q(U)$. We denote a normal vector to $Q(U)$ at a point q by N_q. The *normal vector field* is a mapping $N\colon Q(U) \to \mathbb{R}^n$.

To gain some intuition, consider the dimension $n = 3$, where $Q(U)$ is a two-dimensional surface in $\mathbb{R}^3$. The tangent space $T_qQ(U)$ is then a two-dimensional plane. If v_1 and v_2 are two linearly independent vectors in $T_qQ(U)$ then obviously their cross product is a normal vector to $Q(U)$ at the point q. This idea can be generalized to n-dimensions. Since the cross product can be seen as a determinant, in higher dimensions we can extend this determinental formula to compute a normal vector.

If $\{v_1, \ldots, v_{n-1}\}$ is a set of $n-1$ linearly independent vectors lying in the tangent space $T_qQ(U)$ then,

$$N_q = \begin{vmatrix} e_1 & e_2 & \cdots & e_{n-1} & e_n \\ v_{11} & v_{12} & \cdots & v_{1,n-1} & v_{1n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ v_{n-1,1} & v_{n-1,2} & \cdots & v_{n-1,n-1} & v_{n-1,n} \end{vmatrix}. \tag{3.3}$$

For those familiar with the language of differential geometry, this vector N_q is exactly the Hodge star of the wedge product of the set of vectors $\{v_1, \ldots, v_{n-1}\} \subset T_qQ(U)$, namely

$$N_q = *(v_1 \wedge \cdots \wedge v_{n-1}). \tag{3.4}$$

For an $(n-1) \times n$ matrix A, it is easily seen that

$$\begin{vmatrix} e_1 & \cdots & e_n \\ & & \\ & A & \\ & & \end{vmatrix} \cdot \left(\sum_{j=1}^{n} b_j e_j \right) = \begin{vmatrix} b_1 & \cdots & b_n \\ & & \\ & A & \\ & & \end{vmatrix} . \tag{3.5}$$

From equation (3.5), it is obvious that $N_q \cdot v_i = 0$ for all $i = 1, \ldots, n-1$, proving that N_q calculated by (3.4) is indeed a normal vector.

Recall from Example 3.5 that the partial derivatives $Q_{z_j}|_p$ for p such that $Q(p) = q$, are vectors lying in the tangent space $T_qQ(U)$. We have also noted that, if the Jacobian of our mapping $Q\colon \mathbb{R}^{n-1} \to \mathbb{R}^n$ is one-to-one, the vectors $Q_{z_j}|_p$ are linearly independent in $T_qQ(U)$. Hence to calculate N_q we can simply take the Hodge star of the wedge product of the tangent vectors $Q_{z_j}|_p$, $j = 1, \ldots, n-1$ where $Q(p) = q$. That is,

$$N_q = *(Q_{z_1}|_p \wedge \cdots \wedge Q_{z_{n-1}}|_p). \tag{3.6}$$

Denote by M_q the vector obtained by normalizing N_q to have unit magnitude.

We investigate the shape of $Q(U)$ by observing how the unit normal vector M to $Q(U)$ changes as one moves about the $(n-1)$-dimensional surface $Q(U)$. The *shape operator* or *Weingarten map* does precisely this. The shape operator evaluated at a vector $v \in T_qQ(U)$ is the directional derivative of the unit normal vector M_q in the direction of v. Therefore, the shape operator for a vector v in $T_qQ(U)$ observes how the unit normal vector changes as one moves along $Q(U)$ in the direction of v.

Definition 3.11. The shape operator $L_q\colon T_qQ(U) \to T_qQ(U)$ is defined as

$$L_q(v) = -\nabla_v M = -\nabla M|_q \cdot v.$$

It is easy to see that the linear map L_q maps back to the tangent space $T_qQ(U)$. Indeed, using the product rule of differentiation and the observation that vectors in

$T_qQ(U)$ are perpendicular to M_q, at any point q, we have

$$2(\nabla_v M \cdot M_q) = \nabla_v(M \cdot M) = \nabla_v(1) = 0$$

and hence $\nabla_v M$ is perpendicular to M_q and lies in the tangent space $T_qQ(U)$. It can also be shown that the shape operator is self-adjoint: $L_q(v_1) \cdot v_2 = L_q(v_2) \cdot v_1$ for any two vectors v_1, $v_2 \in T_qQ(U)$.

Using the chain rule, as in expression (3.2), we see that the shape operator can be expressed in terms of any parameterized curve $\alpha(t)$ in $Q(U)$ such that $\alpha(t_0) = q$ and $\alpha'(t_0) = v$. For all such $\alpha(t)$ we have

$$L_q(v) = -\nabla_v M = -(M \circ \alpha)'(t_0). \tag{3.7}$$

The quadratic form $L_q(v) \cdot v$ is called the *second fundamental form.* Theorem 3.12, below indicates that the second fundamental form determines the normal component of acceleration for all curves $\alpha(t)$ embedded in $Q(U)$. In other words, the normal acceleration is imposed on curves by the shape of $Q(U)$.

Theorem 3.12. *Consider $U \subset \mathbb{R}^{n-1}$ and the injective mapping $Q\colon \mathbb{R}^{n-1} \longrightarrow \mathbb{R}^n$ with a one-to-one Jacobian. Then $Q(U)$ is an $(n-1)$-dimensional surface in $\mathbb{R}^n$ with a unit normal vector field M. Let $q \in Q(U)$ and $v \in T_qQ(U)$. For every parametrized curve, $\alpha\colon I \longrightarrow Q(U)$, with $\alpha'(t_0) = v$ and $\alpha(t_0) = q$, for some $t_0 \in I$,*

$$\alpha''(t_0) \cdot M_q = L_q(v) \cdot v.$$

Proof. Obviously $\alpha'(t) \in T_{\alpha(t)}Q(U)$, and is perpendicular to $M_{\alpha(t)}$ for all $t \in I$. Hence $\alpha'(t) \cdot (M \circ \alpha)(t) = 0$.

$$\begin{aligned} 0 &= [\alpha' \cdot (M \circ \alpha)]'(t_0) \\ &= \alpha''(t_0) \cdot (M \circ \alpha)(t_0) + \alpha'(t_0) \cdot (M \circ \alpha)'(t_0) \\ &= \alpha''(t_0) \cdot M(\alpha(t_0)) + v \cdot \nabla_v M \\ &= \alpha''(t_0) \cdot M_q - v \cdot L_q(v) \end{aligned}$$

The proof is now complete. □

When $||v|| = 1$, we set $\kappa(v) = L_q(v) \cdot v$, and call $\kappa(v)$ the *normal curvature* of $Q(U)$ at q in direction v. If $\kappa(v) > 0$ then the surface $Q(U)$ bends towards M_q in the direction v, and if $\kappa(v) < 0$ then the surface $Q(U)$ bends away from M_q in the direction v.

Next, recall a well-known result from linear algebra.

Theorem 3.13. *Let W be a finite-dimensional vector space with dot product and let $L\colon W \longrightarrow W$ be a self-adjoint linear transformation on W. Then there exists an orthonormal basis for W consisting of eigenvectors of L.*

From Theorem 3.13, we know that, in the $(n-1)$-dimensional tangent space $T_qQ(U)$, there exists an orthonormal basis $\{w_1, \ldots, w_{n-1}\}$ which are eigenvectors of the shape operator L_q. We label the corresponding eigenvalues as $k_1(q), \ldots, k_{n-1}(q)$ and call them the *principal curvatures* of $Q(U)$. The corresponding unit eigenvectors are the *principal curvatures directions* of $Q(U)$ at q. If the principal curvatures are ordered such that $k_1(q) \leq \ldots \leq k_{n-1}(q)$, then $k_{n-1}(q)$ is the maximum value of the normal curvature $\kappa(v)$ for $v \in T_qQ(U)$, $||v|| = 1$; $k_{n-2}(q)$ is the maximum value of the normal curvature $\kappa(v)$ for $v \in T_qQ(U)$, $||v|| = 1$, and $v \perp w_{n-1}$ where w_{n-1} is the principal curvature direction corresponding to $k_{n-1}(q)$; $k_{n-3}(q) = \max\{\kappa(v) \mid v \in T_qQ(U), ||v|| = 1, v \perp \{w_{n-1}, w_{n-2}\}\}$, etc. Finally, $k_1(q)$ will be the minimum value of $\kappa(v)$ for $v \in T_qQ(U)$, $||v|| = 1$.

In Example 3.6 we show, using Theorem 3.12, how the curves presented in Example 3.5 can be used to calculate the second fundamental form.

Example 3.6. Second Derivative of Parametrized Curve

Consider again the curve $\alpha(t) = Q(p_1, \ldots, p_{j-1}, t, p_{j+1}, \ldots, p_{n-1})$ parametrized by t and $p \in U$ such that $q = Q(p)$. The second derivative of α at $t = p_j$ is $Q_{z_j,z_j}|_p$.

Therefore the second fundamental form $L_q(v) \cdot v$ where $v = \alpha'(t_0)$ and $\alpha(t_0) = q$ is calculated as $M_q \cdot \alpha''(p_j) = M_q \cdot Q_{z_j z_j}|_p$.

In fact, more generally, for $p \in U$, $Q(p) = q$, and $i \neq j$,

$$L_q(Q_{z_i}|_p) \cdot Q_{z_j}|_p = M_q \cdot Q_{z_i z_j}|_p. \tag{3.8}$$

Indeed, at any point in $Q(U)$, M and Q_{z_i} are perpendicular, and hence

$$\begin{aligned} 0 &= (M \cdot Q_{z_i})_{z_j} \\ &= M_{z_j} \cdot Q_{z_i} + M \cdot Q_{z_i z_j} \end{aligned}$$

We have $-M_{z_j} \cdot Q_{z_i} = M \cdot Q_{z_i z_j}$. To see that $-M_{z_j} = \nabla_{Q_{z_j}} M$, we use the chain rule representation of the directional derivative. Let $\alpha(t) = Q(p_1, \ldots, p_{j-1}, t, p_{j+1}, \ldots, p_{n-1})$. Then $-\nabla_{Q_{z_j}|_p} M = (M \circ \alpha)'(p_j) = M_{z_j}|_p$. In the above, we commit a slight abuse of notation because we treat M as both a function over U and as a function of $Q(U)$, as we have done throughout this section.

3.3 Exponential Families

In Chapter 7, we provide two examples of single-path changepoint problems. One of these examples is based on data distributed as a NEF. Some of the more common NEFs are found in the Morris class of distributions (see Morris (1982) and Morris (1983)). Included in this class are the normal, Poisson, gamma, binomial, and negative binomial distributions. We review NEFs in Section 3.3.1. In Section 3.3.2 we introduce general exponential family (GEF) distributions. The DY-conjugate prior distributions to the NEFs, introduced in Section 3.3.3, are GEFs. An excellent introduction to exponential families is given in a set of lecture notes by Letac (1992). The book by Barndorff-Nielsen (1978) is older and more detailed. The book by Jorgensen (1997) on exponential dispersion models contains a good introduction to

NEFs. Gutiérrez-Pena and Smith (1997) provide an excellent review article concerning conjugate priors for exponential families. In what follows we do not differentiate between a random variable and its realized value. Both random variables and realized values are denoted by lower case letters and whether a quantity is a random variable or a realized variable is evident from the context.

3.3.1 Natural Exponential Families

Let η be a non-degenerate, σ-finite measure on $\mathbb{R}$. Set $\Theta = \{\theta \mid \int e^{\theta x} d\eta(x) < \infty\}$. It is easily shown using Hölder's inequality that Θ is a convex set. In our one-dimensional setting this means that Θ is an interval.

Define the *cumulant transform* $K_\eta(\theta)$ to be the log of the Laplace transform, that is $K_\eta(\theta) = \log\{\int e^{\theta x} d\eta(x)\}$. The family of probabilities indexed by $\theta \in \Theta$ for a single random variable x, with members

$$dP_\theta(x) = \exp\{\theta x - K_\eta(\theta)\} d\eta(x) \tag{3.9}$$

is called a *natural exponential family* (NEF), and Θ is called the *canonical parameter space.* If Θ equals its interior set $int(\Theta)$, then the NEF is said to be *regular.* In the one-dimensional case, this means that a regular family has Θ as an open interval. Note that the *support* of an exponential family is, by definition, the same for all members of the family.

Often the measure η is such that $d\eta(x) = \eta(x) d\eta_o(x)$, where $\eta(x)$ is some non-negative measurable function, and $d\eta_o(x)$ is either the Lebesgue or counting measure. The measure $d\eta(x)$ is then called the *carrier measure.*

Since η is non-degenerate, $K_\eta(\theta)$ is a strictly convex function of θ. It is easily shown that $K'_\eta(\theta) = E(x) = \mu$ and $K''_\eta(\theta) = Var(x)$. We define the *mean domain mapping* as $\tau(\theta) = K'_\eta(\theta)$. Obviously, $\tau(\theta)$ is an increasing function since $\tau'(\theta) = K''_\eta(\theta)$ and $K_\eta(\theta)$ is strictly convex. Hence, the inverse of $\tau(\theta) = \mu$ exists and we have a one-to-one mapping, $\tau^{-1}(\mu) = \theta$, between the mean μ and the canonical parameter θ. The NEF (3.9) can be re-parametrized in terms of the mean μ, and the variance can be expressed as a function of the mean. In fact, the functional relationship between the variance and the mean determines the particular distribution amongst natural exponential families. We call the image $\Omega = \tau(int(\Theta))$ the *mean domain.* If the convex hull of the support is equal to the mean domain Ω, we say the family is *steep.*

Example 3.7. Normal Family with Unknown μ and Fixed σ^2
Consider the normal density,

$$N(x|\mu,\sigma^2) = (2\pi\sigma^2)^{(-1/2)} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

which can be re-written as

$$N(x|\mu,\sigma^2) = \exp\left(\frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2}\right)(2\pi\sigma^2)^{(-1/2)} \exp\left(-\frac{x^2}{2\sigma^2}\right). \qquad (3.10)$$

Equation (3.10) expresses the normal density in its NEF form, from which we identify $\theta = \frac{\mu}{\sigma^2}$, $K_\eta(\theta) = \frac{1}{2}\sigma^2\theta^2$ *and* $\eta(x) = (2\pi\sigma^2)^{(-1/2)} \exp(-\frac{x^2}{2\sigma^2})$.

3.3.2 General Exponential Families

The one-dimensional NEF discussed in Section 3.3.1 can be extended by introducing a vector function of the scalar random variable x. We call this function $t(x)$. The canonical parameter is a vector of the same dimension as the function t. This is a *general exponential family* (GEF) and has the following form.

$$dP_\theta(x) = \exp\{\theta \cdot t(x) - K_\eta(\theta)\}d\eta(x) \qquad (3.11)$$

It is easily shown that $E(t_i(x)) = \frac{\partial K_\eta(\theta)}{\partial \theta_i}$ where $t_i(x)$ is the i^{th} component of $t(x)$ and θ_i is the i^{th} component of θ.

Example 3.8. Normal Family with Unknown μ and σ^2

The normal density function of a random variable x can be expressed as

$$N(x|\mu,\sigma^2) = \exp\left(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2\right)\exp\left(-\frac{\mu^2}{2\sigma^2}\right)(2\pi\sigma^2)^{(-1/2)}. \tag{3.12}$$

With unknown mean and variance, the normal distribution is a GEF. The function $t(x)$ is equal to $(x, -\frac{x^2}{2})$. The canonical parameter is two-dimensional. We identify $(\theta_1, \theta_2) = (\frac{\mu}{\sigma^2}, -\frac{1}{\sigma^2})$, $K_\eta(\theta_1,\theta_2) = \frac{\theta_1}{\theta_2} - \frac{1}{2}\log(\theta_2)$ and $\eta(x) = \frac{1}{\sqrt{2\pi}}$.

3.3.3 Conjugate Priors for NEFs

An NEF has an infinite number of conjugate prior distributions, where conjugate is taken to mean closed under sampling. As Diaconis and Ylvisaker (1979) point out, if $f(\theta)$ is a prior distribution conjugate to $f(x|\theta)$, then, if $h(x)$ is any positive bounded measurable function, $f(\theta)h(x)$ is also a prior distribution conjugate to $f(x|\theta)$. In this thesis we use the *DY-conjugate prior distribution* for the canonical parameter θ of an NEF (3.9). These prior distributions have the form given in (3.13) below, and are characterized by a condition of linearity in x of the posterior expectation $E(K'(\theta)|x)$.

$$dP_{\nu,\lambda}(\theta) = M(\nu,\lambda)\exp\{\nu\theta - \lambda K_\eta(\theta)\}d\theta \tag{3.13}$$

The DY-conjugate prior distributions given in (3.13) were introduced by Diaconis and Ylvisaker (1979) and named after the aforementioned authors. Included in this set of conjugate prior distributions are the normal prior for the mean of the normal distribution, the gamma prior for the Poisson distribution, and the beta prior for the negative binomial distribution. Despite the standard examples just mentioned, the DY-conjugate prior distributions do not always correspond to prior distributions typically used in practice. (See Consonni and Veronese (1992) and Gutiérrez-Pena and Smith (1995) for details.)

Comparing expression (3.13) to expression (3.11), we see that the DY-conjugate prior distribution is actually a GEF. In expression (3.13) the random variable is θ, the vector function t of θ is $(\theta, K_\eta(\theta))$, the canonical parameter is (λ, ν) and $M(\nu, \lambda)$ is a normalizing constant equal to $\frac{1}{\int_\Theta \exp\{\nu\theta - \lambda K_\eta(\theta)\}d\theta}$. The cumulant transform is equal to $-\log(M(\nu, \lambda)) = \log\left(\int_\Theta \exp\{\nu\theta - \lambda K_\eta(\theta)\}d\theta\right)$. There is no carrier measure, as the prior distribution (3.13) is generated by Lebesgue measure. Denoting the cumulant transform as $m(\nu, \lambda)$, we can rewrite (3.13) as

$$dP_{\nu,\lambda}(\theta) = \exp\{\nu\theta - \lambda K_\eta(\theta) - m(\nu, \lambda)\}d\theta. \qquad (3.14)$$

Often λ is referred to as the prior sample size. When (3.13), the DY-conjugate prior for (3.9), is re-expressed in the form of (3.14) we see that $E(K_\eta(\theta)) = \frac{\partial m(\nu,\lambda)}{\partial \lambda}$ and $Var(K_\eta(\theta)) = \frac{\partial^2 m(\nu,\lambda)}{\partial \lambda^2}$.

Example 3.9. Standard Conjugate Prior Distribution for the Canonical Parameter of the Normal NEF

In this case the DY-conjugate prior distribution is equivalent to the usual normal prior distribution for μ. (See Consonni and Veronese (1992) and Gutiérrez-Pena and Smith (1995).) This is easily seen using a transformation of variables from θ to μ. Recall from Example 3.7 that $\theta = \frac{\mu}{\sigma^2}$ for the normal NEF.

The conjugate prior distribution for μ is a normal distribution with mean $\bar{\mu}$ and variance $\bar{\sigma}^2$. From Example 3.8 we express this prior as

$$N(\mu|\bar{\mu}, \bar{\sigma}^2) = \exp\left(\frac{\bar{\mu}}{\bar{\sigma}^2}\mu - \frac{1}{2\bar{\sigma}^2}\mu^2\right)\exp\left(-\frac{\bar{\mu}^2}{2\bar{\sigma}^2}\right)(2\pi\bar{\sigma}^2)^{(-1/2)}. \qquad (3.15)$$

The normal prior distribution (3.15) is presented as a GEF of the random variable μ. The vector function $t(\mu)$ is equal to $(\mu, -\frac{\mu^2}{2})$ and the canonical parameter is $(\frac{\bar{\mu}}{\bar{\sigma}^2}, \frac{1}{\bar{\sigma}^2})$.

Re-parametrizing in terms of θ and $K_\eta(\theta)$ we have

$$\begin{aligned} N(\mu(\theta)|\bar{\mu}, \bar{\sigma}^2) &= \exp\left(\frac{\bar{\mu}\sigma^2}{\bar{\sigma}^2}\left(\frac{\mu}{\sigma^2}\right) - \frac{\sigma^2}{\bar{\sigma}^2}\left(\frac{\sigma^2}{2}\frac{\mu^2}{\sigma^4}\right)\right)\exp\left(-\frac{\bar{\mu}^2}{2\bar{\sigma}^2}\right)(2\pi\bar{\sigma}^2)^{(-1/2)} \\ &= \exp\left(\frac{\bar{\mu}\sigma^2}{\bar{\sigma}^2}\theta - \frac{\sigma^2}{\bar{\sigma}^2}K_\eta(\theta)\right)\exp\left(-\frac{\bar{\mu}^2}{2\bar{\sigma}^2}\right)(2\pi\bar{\sigma}^2)^{(-1/2)}. \end{aligned}$$

This is a GEF in θ, and we identify ν and λ in (3.13) to be $\frac{\bar{\mu}\sigma^2}{\bar{\sigma}^2}$ and $\frac{\sigma^2}{\bar{\sigma}^2}$, respectively.

Chapter 4

The Design Measure

This chapter introduces the design measure and the mapping from the set of allowable designs to the design measure. This is a crucial chapter of the thesis, as the design criterion functions we consider in Chapters 6, 7 and 8 are re-expressed in terms of this design measure. It is important to understand the mapping so that, after optimizing the design criterion functions in terms of the design measure, we can then locate the corresponding optimal design. In Chapter 5, we use the mapping to help us determine the properties of the shape over which we are optimizing the design criterion functions. Here, we introduce the design measure and mapping through the single-path problem.

4.1 The Single-Path Changepoint Model

Our Bayesian model for the single-path changepoint problem consists of

- a likelihood function for n observations $y = (y_1, \ldots, y_n)$;
- design points $x = (x_1, \ldots, x_n)$ at which the observations are taken; and
- prior distributions for the before-and-after-change means and the changepoint.

The vector y denotes our data and the vector x denotes the design. The before-and-after-change means are represented by μ_1 and μ_2, respectively. We restrict ourselves to designs taking measurements a minimum distance d apart, and we assume in our likelihood function that d is large enough to ensure that, roughly, our n measurements $y = (y_1, \ldots, y_n)$ are conditionally independent given μ_1, μ_2, and τ. The likelihood just described is expressed as

$$f(y|\mu_1, \mu_2, \tau) = \prod_{x_i \leq \tau} f(y_i|\mu_1) \prod_{x_i > \tau} f(y_i|\mu_2). \tag{4.1}$$

In fact, we can allow certain observations to be correlated as long as the correlation does not depend on the design; see Chapter 6. However, this does not seem to be a very useful generalization, as in most modelling situations the correlation would usually decrease with distance; of course, if our observations are at least a distance d apart and roughly independent their correlation is constant (in fact, zero). We will see in Section 7.2 that design independent correlation arises and can be accommodated in the common changepoint multi-path problem.

To complete the model, we incorporate the experimenter's uncertainty about μ_1, μ_2 and τ, which is expressed through the prior distribution $f(\mu_1, \mu_2, \tau)$. We shall assume that $f(\mu_1, \mu_2, \tau) = f(\mu_1, \mu_2)f(\tau)$.

4.2 The Design Space

Before introducing the design measure, we describe the set of designs over which we shall optimize. The set of designs consists of all the design vectors $x = (x_1, \ldots, x_n)$ such that $x \in [0, T]^n$ and $0 \leq x_1$, $x_{i-1} + d < x_i$ for $i = 2, \ldots, n$ and $x_n \leq T$. For each experiment, d is selected so that observations a distance d or more apart can assumed to be conditionally independent given μ_1, μ_2, and τ. We denote the set of all possible designs by $\mathbb{X}^n$, where n reminds us that n observations are taken. We refer to the set of all possible designs $\mathbb{X}^n$ as the *design space.*

In Theorem 4.1 we prove that $\mathbb{X}^n$ forms a simplex. Let V be the vertex set of $\mathbb{X}^n$. The designs in V are $(0, d, \ldots, (n-1)d)$ which places all observations towards 0; $(0, d, \ldots, (n-2)d, T)$ which places $n-1$ observations towards 0 and the nth observation at T, through to the design $(T-(n-1)d, T-(n-2)d, \ldots, T)$ which places all the observations towards T. There are $n+1$ such designs in V and we label them u_0 to u_n, respectively. Thus, u_i indicates that i design points are placed towards T. So

$$V = \{u_0, \ldots, u_n\}.$$

Note that for all our experiments we assume that $(n-1)d < T$, so that the n observations fit into the observation interval $[0, T]$ with the constraint that they are all a minimum distance d apart.

Theorem 4.1. *The design space $\mathbb{X}^n$, where $x \in \mathbb{X}^n$ implies that $x \in [0, T]^n$ and $0 \leq x_1$, $x_{i-1} + d < x_i$ for $i = 2, \ldots, n$ and $x_n \leq T$, forms an n-dimensional simplex whose vertices correspond to the designs in the set V placing points as far as possible towards the ends of the observation interval* $[0, T]$.

Proof. First we show that V is an affinely independent set. Using Corollary 3.2 from Section 3.1, we prove that the set $\{u_0, \ldots, u_n\}$ is affinely independent by proving that the set $\{u_1 - u_0, \ldots, u_n - u_0\}$ is linearly independent. A well-known result from linear algebra states that if the determinant of n vectors in n-dimensional space is non-zero then the vectors are linearly independent. Hence, we consider the following determinant of the set of vectors $\{u_1 - u_0, \ldots, u_n - u_0\}$.

$$\begin{vmatrix} 0 & 0 & \ldots & 0 & T-(n-1)d \\ 0 & 0 & \ldots & T-(n-1)d & T-(n-1)d \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ T-(n-1)d & T-(n-1)d & \ldots & T-(n-1)d & T-(n-1)d \end{vmatrix}$$

Since this is the determinant of a matrix with an upper triangle of zeros, the determinant is proportional to the product of the diagonal entries. In this case the determinant is $(-1)^{\lfloor \frac{n}{2} \rfloor}(T-(n-1)d)^n$ which is non-zero.

It remains to show that every design lies in the convex hull of V. That is, we wish to show that $x \in \mathbb{X}^n$ can be written as $x = \sum_{i=0}^{n} \lambda_i u_i$, where $\lambda_i \geq 0$ for all $i = 0, \ldots, n$ and $\sum_{i=0}^{n} \lambda_i = 1$. This is easily seen to be the case by substituting $\lambda_n = \frac{x_1}{T-(n-1)d}$, $\lambda_{n-1} = \frac{x_2 - x_1 - d}{T-(n-1)d}$, $\ldots$, $\lambda_1 = \frac{x_n - x_{n-1} - d}{T-(n-1)d}$, and $\lambda_0 = \frac{T - x_n}{T-(n-1)d}$ into the affine combination $\sum_{i=0}^{n} \lambda_i u_i$. The constraints satisfied by $x \in \mathbb{X}^n$ ensure that $\lambda_i > 0$ for all $i = 0, \ldots, n$ and that $\sum_{i=0}^{n} \lambda_i = 1$. Therefore we have proved that the design space $\mathbb{X}^n$ is a simplex. □

Next we introduce the design measure used throughout this thesis.

4.3 The Design Measure

A design measure appears quite naturally in the changepoint problem. It is the discrete probability measure of a random variable which is formed from the changepoint random variable τ and the deterministic design vector x. Consequently, we define the new random variable τ_x as follows: for $i = 0, 1, \ldots, n$ and setting x_0 and x_{n+1} to be 0 and T, respectively, let

$$I_{x_i}(\tau) = \begin{cases} i, & \text{if } x_i \leq \tau < x_i + 1 \\ 0, & \text{otherwise.} \end{cases} \tag{4.2}$$

In the case of point mass at T,

$$I_{x_n}(\tau) = \begin{cases} n, & \text{if } x_n \leq \tau \leq T \\ 0, & \text{otherwise.} \end{cases} \tag{4.3}$$

Let

$$\tau_x = \sum_{i=0}^{n} I_{x_i}(\tau).$$

So τ_x is the number of design points x_i in the design x which satisfy $x_i \leq \tau$. The event $\{\tau_x = k\}$ is defined as the event that the change occurs at x_k or between the design points x_k and x_{k+1}. This event is equivalent to the event that the first k observations $(y_1, \ldots, y_k)$ come from the distribution $f(y|\mu_1)$ and the last $n-k$ observations $(y_{k+1}, \ldots, y_n)$ come from the distribution $f(y|\mu_2)$. Since there are n design points in the observation interval $[0, T]$ there are $n+1$ intervals in which the changepoint can fall. As we only consider prior distributions for τ that are continuous on $[0, T)$, the distribution of τ_x has the $n+1$ support points $0, 1, \ldots, n$ with probabilities

$$\pi_k \overset{\text{def}}{=} P(\tau_x = k) = P(x_k \leq \tau < x_{k+1}) = \int_{x_k}^{x_{k+1}} f, \tag{4.4}$$

where we ignore the dummy variable of integration. In the case of a probability mass at T in our changepoint prior density we have

$$\pi_n \overset{\text{def}}{=} P(\tau_x = n) = P(x_n \leq \tau). \tag{4.5}$$

The vector $\pi = (\pi_0, \ldots, \pi_n)$ is our design measure. Once the prior distribution for the changepoint has been selected, we consider it to be fixed and the design measure π is then a function of only the design x. Figure 4.1 illustrates the situation when $n = 2$. Obviously the areas π_0, π_1, and π_2 change as the design points x_1 and x_2 move location in the observation interval $[0, T]$.

As mentioned in Section 3.1, we represent the set of all design measures by the barycentric coordinates of the specific simplex S^n. Since we restrict our design points to be at least a distance d apart, we minimize our design criterion functions over a subset of S^n. The subset over which we minimize depends on the prior density $f(\tau)$. To better understand this, we examine the mapping from the design x to the design measure π.

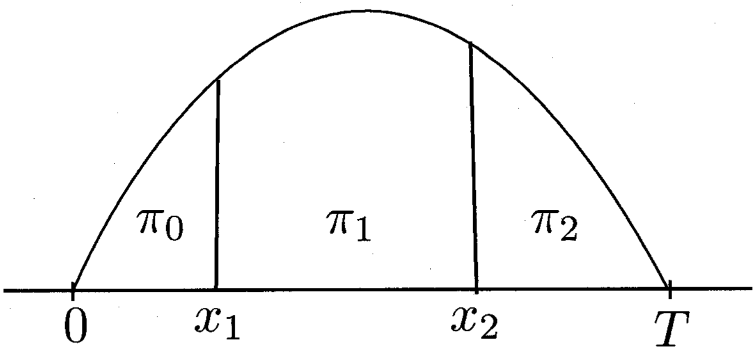


Figure 4.1: A unimodal prior distribution $f(\tau)$ on the interval $[0, T]$ and a two-dimensional design $x = (x_1, x_2)$. The component areas, π_0, π_1, and π_2 of the design measure are also shown.

4.4 The Mapping from the Design x to the Design Measure π

Define a mapping $G_f \colon \mathbb{X}^n \to S^n$ as $G_f(x_1, \ldots, x_n) = (\pi_0, \ldots, \pi_n)$. The subscript, f, serves as a reminder that the mapping depends on the prior distribution of the changepoint τ.

Equation (4.4) gave the mapping for the kth element of π. The complete mapping is as follows,

$$G_f(x) = (\pi_0(x), \ldots, \pi_n(x)) = \left(\int_0^{x_1} f, \int_{x_1}^{x_2} f, \ldots, \int_{x_n}^{T} f \right). \tag{4.6}$$

To express (4.6) in terms of the Euclidean coordinates z of S^n, we simply drop the first component π_0, hence selecting the affinely independent set $\{e_0, \ldots, e_n\}$ as the vertices of S^n. Therefore, as discussed in Section 3.1, we have the unique relationship

$\pi_0 = 1 - z_1 - \cdots - z_n$, $\pi_1 = z_1, \ldots, \pi_n = z_n$. In terms of the Euclidean coordinates, we now define

$$G_f(x) = (\pi_1(x), \ldots, \pi_n(x)) = \left(\int_{x_1}^{x_2} f, \int_{x_2}^{x_3} f, \ldots, \int_{x_n}^{T} f \right). \tag{4.7}$$

This mapping is injective if the prior density f equals zero only on sets of Lebesgue measure zero; if f equals zero on sets of measure greater than zero then there could be multiple designs mapping to the same design measure.

Consideration of the *design measure*, rather than the *design*, allows us to focus on the essence of the structure. What is important is the probability of the change occurring between the design points rather than the explicit distances between the design points. Our interest lies in determining the subset $G_f(\mathbb{X}^n)$ over which we minimize the design criterion functions. We stress that the shape of this subset depends on the prior density f for the changepoint τ.

Chapter 5

The Shape of $G_f(\mathbb{X}^n)$

In this chapter we investigate how the prior distribution for the changepoint affects the shape over which we minimize our various design criterion functions. Initially our design criterion functions are functions of the design x and the prior f on τ. In Chapters 6, 7, and 8 we combine the design x and the prior density f and re-write the design criterion functions in terms of π. As functions of π, the design criterion functions are concave functions and are much easier to work with than the original design criterion functions over $\mathbb{X}^n$. Although the criterion functions become more tractable, the shapes over which we minimize do not. When the criterion functions are functions of the design x, we minimize over the set of all possible designs $\mathbb{X}^n$, which is a simplex. Now, as functions of π, we minimize over $G_f(\mathbb{X}^n)$. By Theorem 3.8, if $G_f(\mathbb{X}^n)$ is a simplex we could easily minimize the concave criterion functions. However, for most prior densities, f, the set $G_f(\mathbb{X}^n)$ is not a simplex. Corollary 3.1, though, asserts that if a concave function is minimized over a set which is a subset of a simplex and contains the vertices of that simplex then the minimum must occur at one of the vertices of the simplex. Here, we find a set of priors f such that $G_f(\mathbb{X}^n)$ is either a simplex or is a subset of a simplex containing the vertices of the simplex. While proving that $G_f(\mathbb{X}^n)$ is the subset of a simplex, we also prove that the

$n + 1$ designs that place points as far as possible towards the end of the observation interval $[0, T]$ (that is, the vertex set V) are the designs which map to the vertices of the simplex associated with G_f. This is an important result because after we find the optimal design measure (the measure π which minimizes the design criterion function) we need to find the optimal design (the design x corresponding to the optimal design measure π). The optimal design is the design which minimizes the design criterion function, *as a function of the design.*

5.1 Examples and Motivation

Before considering the n-dimensional problem with n design points, for illustration, we first consider some simple examples with two design points.

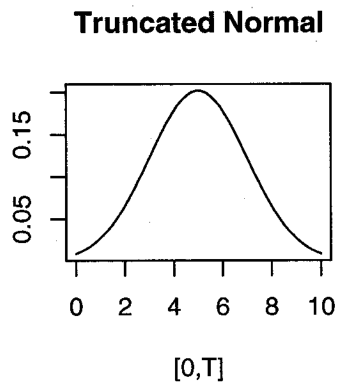


Figure 5.1: A truncated normal prior density, truncated between 0 and T centered at $\frac{T}{2}$ with standard deviation 2. Here T is equal to 10.

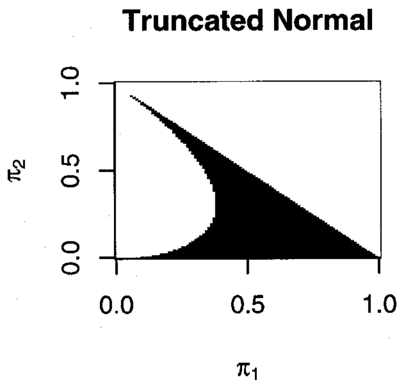


Figure 5.2: The set $G_f(\mathbb{X}^2)$ with d equal to 2 for the truncated normal prior in Figure 5.1.

Example 5.1. Truncated Normal

Figure 5.1 shows a unimodal prior density for τ *and Figure 5.2 shows the correspond-*

ing image $G_f(\mathbb{X}^2)$ with d equal to 2. It is easily seen that $G_f(\mathbb{X}^2)$ contains the vertices of a simplex and is a subset of that simplex.

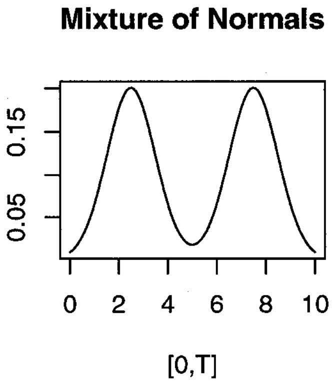


Figure 5.3: A half-and-half mixture of two normal prior densities truncated between 0 and T. The normal components are centered at $\frac{T}{4}$ and $\frac{3T}{4}$, respectively. Their standard deviations are 1, and T is 10.

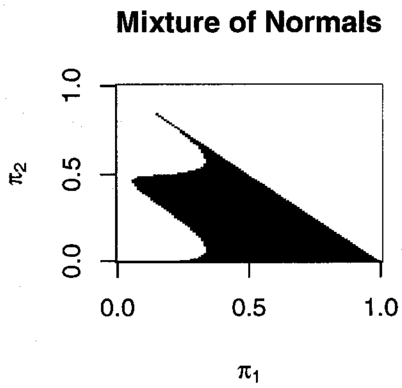


Figure 5.4: The set $G_f(\mathbb{X}^2)$ with d equal to 2 for the mixture of two truncated normal prior densities in Figure 5.3.

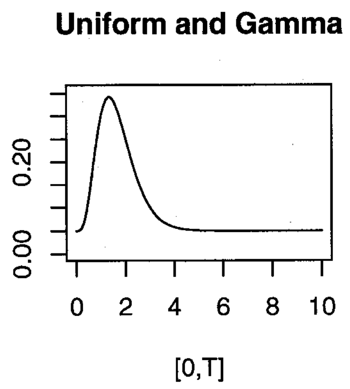
Example 5.2. Mixture of Normals

Figure 5.3 shows a bimodal prior density for τ and Figure 5.4 shows the corresponding image $G_f(\mathbb{X}^2)$ with d equal to 2. Here $G_f(\mathbb{X}^2)$ contains the vertices of a simplex but is not a subset of that simplex.

Examples 5.1 and 5.2 suggest that if f is unimodal then $G_f(\mathbb{X}^n)$ is a subset of a simplex and contains the vertices of that simplex. However, as the next example illustrates, this is not always the case.

Example 5.3. Mixture of a Uniform and a Gamma

Figure 5.5 shows a skewed unimodal prior with heavy tails and Figure 5.6 shows the



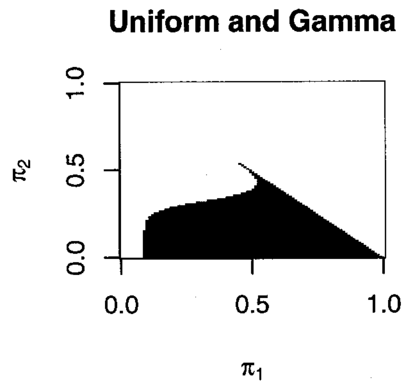


Figure 5.5: A half-and-half mixture of a uniform distribution between 0 to T and a gamma distribution with shape parameter 5 and scale parameter $\frac{1}{3}$. The gamma distribution is truncated between 0 and T, with T equal to 10.

Figure 5.6: The set $G_f(\mathbb{X}^2)$ with d equal to 2 for the mixture of the uniform and gamma prior densities in Figure 5.5.

corresponding space $G_f(\mathbb{X}^2)$ with d equal to 2. Obviously, even though the prior is unimodal, $G_f(\mathbb{X}^2)$ is not a subset of a simplex whose vertices are in $G_f(\mathbb{X}^2)$.

Consequently, Example 5.3 invites the question: Is $G_f(\mathbb{X}^n)$ not a subset of a simplex whose vertices are in $G_f(\mathbb{X}^n)$ because it is skewed or because it has a heavy tail? We show that for any log-concave prior density f, the image $G_f(\mathbb{X}^n)$ lies inside a simplex and contains the vertices of the simplex. In Example 5.3, it is the heavy tail and not the skewness of the prior that violates the log-concavity. In fact, there are many skewed log-concave distributions. As shown in Bagnoli and Bergström (1989) the normal, chi-squared, and extreme-value density functions are all strictly log-concave and the Weibull, power variance function, gamma, and beta densities are log-concave for certain parameter values. They also show that the truncated density of any log-concave density is also log-concave. Therefore, for our optimal design problem, we can use any truncated version of the densities just mentioned (with an appropriate choice of parameters when required) and, in fact, any other log-concave (not necessarily strictly log-concave) density for our prior density $f(\tau)$ on $[0, T]$.

Our task is now to prove that, for every log-concave f, the set $G_f(\mathbb{X}^n)$ lies inside a simplex and includes the vertices of that simplex. For the proof, we restrict our attention to differentiable prior densities f with support in $[0, T]$. Initially, in Section 5.2, we consider prior densities f that are everywhere greater than zero, and then in Section 5.3, we extend the proof to three cases for which f is zero over an interval of Lesbgue measure greater than zero. In Section 5.4, we consider the case when the prior density is differentiable over $[0, T)$ but has mass at T; the mass at T allows for the possibility that there is no changepoint. We conclude in Sections 5.5 and 5.6 by addressing the implications of our log-concave prior density result for design criterion functions concave in π.

For each of the different types of prior densities, our proof consists of two parts. In the first part, we show that the set $G_f(V)$ is affinely independent. It follows that

$\mathrm{Conv}(G_f(V))$, which we denote as $\triangle_f$, is a simplex. In the second part, we consider each $(n-1)$-dimensional boundary of $G_f(\mathbb{X}^n)$ and find conditions on f such that $G_f(\mathbb{X}^n) \subset \triangle_f$.

5.2 Proof for f Everywhere Greater than Zero

5.2.1 Part I: Affine Independence of $G_f(V)$

We use the results presented in Section 3.1 to prove that $G_f(V)$ is affinely independent for a prior density f that is everywhere greater than zero in $[0,T]$. Recall the vertex set $V = \{u_0, \dots, u_n\}$ in (4.2), where u_i places i design points towards T.

Lemma 5.1. *For f everywhere greater than zero (not necessarily continuous or differentiable), the set $G_f(V)$ is affinely independent.*

Proof. Let

$$\alpha_i = \int_{(i-1)d}^{id} f, \quad \beta_i = \int_{(n-i-1)d}^{T-(i-1)d} f, \quad \gamma_i = \int_{T-id}^{T-(i-1)d} f,$$

for $i = 1, \dots, n-1$, and let

$$\bar{\alpha} = \int_0^{T-(n-1)d} f, \quad \bar{\beta} = \int_{(n-1)d}^{T} f.$$

Dropping π_0, and writing out the Euclidean coordinates of the set $G_f(V)$, we have

$$\begin{aligned} G_f(u_0) &= (\alpha_1, \dots, \alpha_{n-1}, \bar{\beta}), \\ G_f(u_1) &= (\alpha_1, \dots, \alpha_{n-2}, \beta_1, 0), \\ &\vdots \\ G_f(u_j) &= (\alpha_1, \dots, \alpha_{n-j-1}, \beta_j, \gamma_{j-1}, \dots, \gamma_1, 0), \\ &\vdots \\ G_f(u_{n-1}) &= (\beta_{n-1}, \gamma_{n-2}, \dots, \gamma_1, 0), \\ G_f(u_n) &= (\gamma_{n-1}, \dots, \gamma_1, 0). \end{aligned}$$

By Corollary 3.2, to prove that $G_f(V)$ is an affinely independent set, we can show that the set of vectors $\{G_f(u_0) - G_f(u_n), G_f(u_1) - G_f(u_n), \ldots, G_f(u_{n-1}) - G_f(u_n)\}$ is linearly independent. This set is linearly independent if the determinant

$$\begin{vmatrix} \alpha_1 - \gamma_{n-1} & \alpha_2 - \gamma_{n-2} & \ldots & \alpha_{n-1} - \gamma_1 & \bar{\beta} \\ \alpha_1 - \gamma_{n-1} & \alpha_2 - \gamma_{n-2} & \ldots & \beta_1 - \gamma_1 & 0 \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ \alpha_1 - \gamma_{n-1} & \beta_{n-2} - \gamma_{n-2} & \ldots & 0 & 0 \\ \beta_{n-1} - \gamma_{n-1} & 0 & \ldots & 0 & 0 \end{vmatrix} \tag{5.1}$$

is non-zero. Obviously, the determinant of a triangular matrix is non-zero if all elements on the diagonal are non-zero. This is indeed the case here. To see this, observe that $\beta_j - \gamma_j = \int_{(n-j-1)d}^{T-jd} f$, and f is everywhere greater than zero. Similarly, $\bar{\beta} = \int_{(n-1)d}^{T} f$ is greater than zero. □

From Definition 3.4, the convex hull of the affinely independent set $G_f(V)$ is a simplex which we denote as $\triangle_f$. Next, we find a set of differentiable f everywhere greater than zero such that $G_f(\mathbb{X}^n) \subset \triangle_f$.

5.2.2 Part II: Condition on f so that $G_f(\mathbb{X}^n) \subset \triangle_f$

In this section we use the results from Section 3.2 to prove that, for a log-concave prior density that is everywhere greater than zero, $G_f(\mathbb{X}^n) \subset \triangle_f$. Before showing this, we introduce some notation used in its proof.

We label the facets of $\mathbb{X}^n$ as F_0 through to F_n, where

$$\begin{aligned} F_0 :& \quad x_1 = 0, \\ F_i :& \quad x_{i+1} = x_i + d, \quad (1 \leq i < n) \\ F_n :& \quad x_n = T. \end{aligned} \tag{5.2}$$

Each F_i is the convex hull of every vertex except the vertex u_{n-i} (which has the largest distance possible between the points x_i and x_{i+1}). The facet F_i is parametrized by the $(n-1)$-dimensional vector $(x_1, \ldots, x_i, \hat{x}_{i+1}, x_{i+2}, \ldots, x_n)$. Hence $G_f(F_i)$ is an $(n-1)$-dimensional surface lying in $\mathbb{R}^n$.

To simplify notation, let $G = G_f$ and let G^i denote the restriction $G_f|_{F^i}$ of G_f to the subset F^i of $\mathbb{X}^n$:

$$G^i(x_1, \ldots, x_i, \hat{x}_{i+1}, x_{i+2}, \ldots, x_n) = G_f(x_1, \ldots, x_i, x_i + d, x_{i+2}, \ldots, x_n).$$

We denote partial derivatives by subscripts. Recall from Section 3.2, if $p \in F^i$ is such that $G^i(p) = q$ then the partial derivatives $G^i_j|_p$, $j \neq i+1$, lie in the tangent space $T_q G^i$. The partial derivatives are

$$\begin{aligned}
G^i_1 &= (-f(x_1), 0, \ldots, 0), \\
G^i_2 &= (f(x_2), -f(x_2), 0, \ldots, 0), \\
&\vdots \\
G^i_{i-1} &= (0, \ldots, f(x_{i-1}), -f(x_{i-1}), 0, \ldots, 0), \\
G^i_i &= (0, \ldots, f(x_i), f(x_i + d) - f(x_i), -f(x_i + d), 0, \ldots, 0), \\
G^i_{i+2} &= (0, \ldots, 0, f(x_{i+2}), -f(x_{i+2}), 0, \ldots, 0), \\
&\vdots \\
G^i_{n-1} &= (0, \ldots, 0, f(x_{n-1}), -f(x_{n-1}), 0), \\
G^i_n &= (0, \ldots, 0, f(x_n), -f(x_n)).
\end{aligned} \tag{5.3}$$

It is easy to show that, since f is a density function always greater than zero, the partial derivatives (5.3) are linearly independent in any tangent space $T_q G^i$.

Using expression (3.6), we calculate a normal vector N^i to $G_f(F_i)$, as the Hodge star of the wedge product of the partial derivatives (5.3):

$$N^i = *(G^i_1 \wedge \cdots \wedge G^i_i \wedge G^i_{i+2} \wedge \cdots \wedge G^i_n). \tag{5.4}$$

So

$$
\begin{aligned}
N^i & \\
= & \begin{vmatrix}
e_1 & e_2 & \cdots & e_{i-1} & e_i & e_{i+1} & \cdots & e_{n+1} & e_n \\
-f(x_1) & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\
f(x_2) & -f(x_2) & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & -f(x_{i-1}) & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & \cdots & f(x_i) & f(x_i+d)-f(x_i) & -f(x_i+d) & \cdots & 0 & 0 \\
0 & 0 & \cdots & 0 & 0 & f(x_{i+2}) & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & f(x_{n-1}) & -f(x_{n-1}) \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & -f(x_n) &
\end{vmatrix} \\
= & \quad \textstyle\sum_{j=1}^{i-1} 0e_j + (-1)^{n-i} f(x_1)\cdots f(x_{i-1})f(x_i+d)f(x_{i+2})\cdots f(x_n)e_i + \\
& \quad \textstyle\sum_{j=i+1}^{n} (-1)^{n-i} f(x_1)\cdots f(x_{i-1})[f(x_i+d)-f(x_i)]f(x_{i+2})\cdots f(x_n)e_j. \qquad (5.5)
\end{aligned}
$$

Obviously, the determinential formula above is slightly different if $i = 1$ or $i = n-1$, but the final formula obtained for N^i is true for $i = 1, \ldots, n-1$.

To determine if N^i is inward or outward pointing we take the dot product of N^i and a vector known to be inward pointing. The vector G_{i+1}, evaluated at any $p \in F_i$, is inward pointing because G_f no longer maps to F_i as x_{i+1} moves away from x_i. With $f(x_{i+1})$ as the ith component and $-f(x_{i+1})$ as the $(i+1)$st component we have

$$G_{i+1} = (0, \ldots, 0, f(x_{i+1}), -f(x_{i+1}), 0, \ldots, 0).$$

Hence

$$N^i \cdot G_{i+1} = (-1)^{n-i} f(x_1)\cdots f(x_{i-1})f(x_i)f(x_{i+1})\cdots f(x_n).$$

Obviously the factor $(-1)^{n-i}$ determines the sign of $N^i \cdot G_{i+1}$ and, hence, whether N^i points inward or outward alternates with i. If $n-i$ is even then N^i is inward pointing; if $n - i$ is odd then N^i is outward pointing. To simplify our work we set

$$\bar{N}^i = (-1)^{n-i} N^i, \tag{5.6}$$

so that $\bar{N}^i$ always points inward.

Quick inspection shows that all mixed partial derivatives are zero. The second partial derivatives G^i_{jj} are

$$\begin{aligned}
G^i_{11} &= (-f'(x_1), 0, \dots, 0), \\
G^i_{22} &= (f'(x_2), -f'(x_2), 0, \dots, 0), \\
&\vdots \\
G^i_{i-1,i-1} &= (0, \dots, f'(x_{i-1}), -f'(x_{i-1}), 0, \dots, 0), \\
G^i_{ii} &= (0, \dots, f'(x_i), f'(x_i + d) - f'(x_i), -f'(x_i + d), 0, \dots, 0), \\
G^i_{i+2,i+2} &= (0, \dots, 0, f'(x_{i+2}), -f'(x_{i+2}), 0, \dots, 0), \\
&\vdots \\
G^i_{n-1,n-1} &= (0, \dots, 0, f'(x_{n-1}), -f'(x_{n-1}), 0), \\
G^i_{nn} &= (0, \dots, 0, f'(x_n), -f'(x_n)).
\end{aligned} \tag{5.7}$$

We are now ready to prove Theorem 5.1.

Theorem 5.1. *If the prior f is log-concave and everywhere greater than zero then $G_f(\mathbb{X}^n) \subset \triangle_f$, where $\triangle_f = \mathrm{Conv}(G_f(V))$.*

Proof. Clearly, the set $G_f(V) \subset G_f(\mathbb{X}^n)$ lies in $\triangle_f$. To determine f such that $G_f(\mathbb{X}^n) \subset \triangle_f$, we examine each of the $n + 1$ boundaries of $\triangle_f$ and find a condition on f to ensure that the corresponding boundaries of $G_f(\mathbb{X}^n)$ lie inside their counterpart boundaries of $\triangle_f$. A boundary of $G_f(\mathbb{X}^n)$ is paired with a boundary of $\triangle_f$ if they both attach to the same subset of n vertex points of $G_f(V)$. Essentially, the paired boundaries are formed from the same facet F_i of $\mathbb{X}^n$. Recall from Section

4.4, since f is everywhere greater than zero, G_f is an injective mapping and, thus, maps boundaries of $\mathbb{X}^n$ to boundaries of $G_f(\mathbb{X}^n)$.

Hence, our interest lies in the boundaries $G_f(F_i)$ parametrized by G^i. Since G^0 parametrizes $\pi_0 = 0$ and G^n parametrizes $\pi_n = 0$, the boundaries parametrized by G^0 and G^n are coincident with their corresponding boundaries in $\triangle_f$. Consequently, we focus on the boundaries parametrized by G^i for $i = 1, \dots, n-1$.

We note that if the principal curvatures calculated according to the inward normal pointing vector $\bar{N}^i$, defined in (5.6), are all non-positive at every point $p \in F_i$, then the shape of $G_f(F_i)$ is such that $G_f(F_i)$ is "pulled" inside $\triangle_f$.

Define the normal vector $\bar{M}^i$ as the normalized $\bar{N}^i$. Our subsequent discussion applies at any point $p \in F_i$. However, for simplicity, we do not make this dependence explicit. Denoting the shape opeator as L, we see from Theorem 3.12 and Example 3.6 that

$$L(G^i_j) \cdot G^i_j = \bar{M}^i \cdot G^i_{jj}$$

and from expression (3.8) that

$$L(G^i_j) \cdot G^i_k = \bar{M}^i \cdot G^i_{jk}.$$

Since all mixed partial derivatives are zero, $L(G^i_j) \cdot G^i_k$ is zero for all j and k. Furthermore, quick calculation shows that $\bar{N}^i \cdot G^i_{jj}$ and hence $L(G^i_j) \cdot G^i_j$ equals zero for all $j \neq i$. Therefore we are left with

$$\begin{aligned} \bar{N}^i \cdot G^i_{ii} &= (-1)^{n-i} f(x_1) \cdots f(x_{i-1}) f(x_{i+2}) \cdots f(x_n) \\ &\quad \times \big((f'(x_i + d) - f'(x_i)) f(x_i + d) - f'(x_i + d)(f(x_i + d) - f(x_i))\big) \end{aligned} \tag{5.8}$$

which determines the sign of $L(G^i_i) \cdot G^i_i$. From (5.8), we see that the expression

$$(f'(x_i + d) - f'(x_i)) f(x_i + d) - f'(x_i + d)(f(x_i + d) - f(x_i)) \tag{5.9}$$

determines the sign of $L(G^i_i) \cdot G^i_i$.

To summarize, for some constant A, we have

$$L(G^i_j) \cdot G^i_k = \begin{cases} A, & \text{when } i = j = k, \\ 0, & \text{otherwise.} \end{cases} \tag{5.10}$$

Since the shape operator L is self adjoint, we know from Theorem 3.13 that there exists an orthonormal basis consisting of eigenvectors $\{w_1, \dots, w_{n-1}\}$ for L. The eigenvalues are the principal curvatures, hence we need to find their signs.

Denote the Kronecker delta function by δ_{jk} and the principal curvature for w_j by k_j. We then have

$$\begin{aligned} w_j \cdot w_k &= \delta_{jk} \\ L(w_j) \cdot w_k &= \delta_{jk} k_j. \end{aligned} \tag{5.11}$$

Our goal is to relate the sign of $L(G^i_i) \cdot G^i_i$ to the signs of the k_j's.

Since $\{G^i_i, \dots, G^i_i, G^i_{i+2}, \dots, G^i_n\}$ and $\{w_1, \dots, w_{n-1}\}$ are two bases of the same space, we find a matrix $[h_{lj}]$ such that $w_l = \sum_k h_{lk} G^i_k$. The principal curvatures or eigenvalues of the shape operator L are

$$\begin{aligned} k_j &= L(w_j) \cdot w_j \\ &= \sum_{k,l} h_{jk} h_{jl} L(G^i_k) \cdot G^i_l \\ &= A(h_{ji})^2. \end{aligned} \tag{5.12}$$

The signs of the principal curvatures are determined by A. As stated, we need our principal curvatures to have non-positive sign. Accordingly, we make the expression (5.9) non-positive. Simplifying, we find the condition

$$f(x_i) f'(x_i + d) - f'(x_i) f(x_i + d) \leq 0,$$

or equivalently that

$$\frac{f'(x+d)}{f(x+d)} \leq \frac{f'(x)}{f(x)}. \tag{5.13}$$

Any prior f with $\frac{f'}{f}$ monotone decreasing satisfies equation (5.13). Differentiating $\log f$ twice we immediately see that $\frac{f'}{f}$ being monotone decreasing is equivalent to f being log-concave. The result follows. $\square$

The converse of Theorem 5.1 does not hold. It is not true that $G_f(\mathbb{X}^n) \subset \triangle_f$ implies that f is log-concave. Our constraint on the principle curvatures was stronger than required; there could be non-log-concave f, such that $G_f(\mathbb{X}^n) \subset \triangle_f$. Nonetheless, the class of log-concave prior densities is sufficiently rich to contain most reasonable prior distributions that might be considered for the changepoint.

5.3 When f Equals Zero Over Part of $[0, T]$

In Section 5.2, we restricted ourselves to prior densities f that were differentiable and everywhere greater than zero. We showed that if f is log-concave, then $G_f(\mathbb{X}^n) \subset \triangle_f$. As we prove our design criterion functions are concave functions of the design measure, we know that for any log-concave prior density, everywhere greater than zero, a concave criterion function is minimized at one of the design measures in $G_f(V)$. Ergo our optimal design is one of the $n+1$ designs in V placing points as far as possible towards the ends of the interval $[0, T]$, while maintaining a distance d between them.

Although Theorem 5.1 is an elegant result, it is worth considering if we could extend the result to include prior densities which are not always greater than zero. For instance, the motivational Example 1.1 in the introduction concerned a changepoint caused by a blood pressure lowering treatment. We argued that, although the time the drug was administered would be known, the time the drug took effect would not be known; hence the unknown changepoint. A prior distribution for the changepoint describes the uncertainty regarding when the treatment would take effect. We are certain, however, that the treatment would not take effect before it was administered. Therefore, if the treatment is administered at some time $t > 0$, we would want a prior density for τ to be zero over the interval $[0, t]$ and everywhere greater than zero in the interval $(t, T]$; this is the Type 1 prior depicted in Figure 5.7. One could also imagine situations where the Type 2 and Type 3 priors in Figure 5.7 would be of interest.

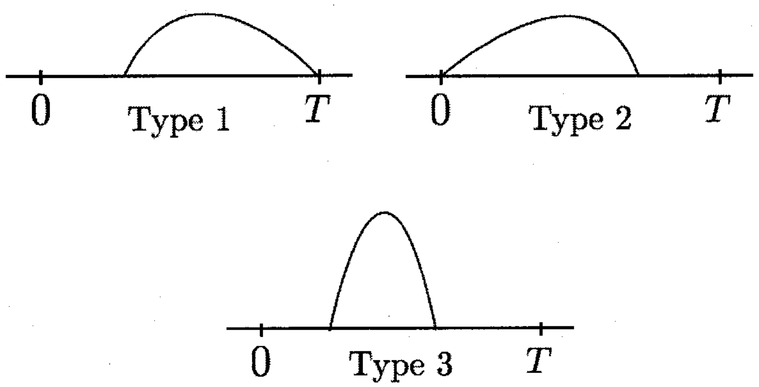


Figure 5.7: Representations of three types of differentiable prior densities zero over an interval in $[0, T]$. The Type 1 prior density is zero at the start of the interval $[0, T]$; the Type 2 prior density is zero at the end of the interval $[0, T]$; the Type 3 prior density is zero both at the start and at the end of the interval $[0, T]$.

We have two difficulties with these three types of prior densities because they equal zero over an interval which has Lebesgue measure greater than zero. The first problem is that the set $G_f(V)$ might not be affinely independent. The second is that the mapping G_f is not injective and the Jacobian is not one-to-one for designs with design points which are not in the support of f. For such designs we are unable to calculate the normal vector used in our proof of log-concavity in Section 5.2. Fortunately, as we shall see there are analogues to Lemma 5.1 and Theorem 5.1 that account for Type 1, Type 2, and Type 3 prior densities, respectively. In these analogous results we find that the dimension of the problem is often reduced from n. We use the notation that $G_{f,(i,j)}$ equals G_f but provides only the components $(\pi_i, \ldots, \pi_j)$.

5.3.1 Type 1 Prior

Lemma 5.2. *For a Type 1 prior density f:*

A) If the support $[t, T]$ is such that $(T - t) > (n-1)d$ then the set $G_f(V)$ is affinely independent.

B) If support $[t, T]$ is such that $(n-k)d > [t, T] > (n-k-1)d$ then the set $G_{f,(k,n)}(\{u_0, \ldots, u_{n-k}\})$ is affinely independent.

Sketch of Proof. Statement A) is easily shown via a proof similar to that of Lemma 5.1. In situation B), where $(n-k)d > (T-t) > (n-k-1)d$, there exists no design in $\mathbb{X}^n$ where the first k design points x_1 to x_k appear in the support of f. Consequently, π_0 to π_{k-1} always equal zero and the set $G_f(V)$ is not affinely independent. However, again using a proof similar to that of Lemma 5.1, we can show the set $G_{f,(k,n)}(\{u_0, \ldots, u_{n-k}\})$ is affinely independent. □

Theorem 5.2. *For a log-concave Type 1 prior density f:*

A) If the support $[t, T]$ is such that $(T-t) > (n-1)d$ then $G_f(\mathbb{X}^n) \subset \triangle_f$, where $\triangle_f = \text{Conv}(G_f(V))$.

B) If the support is such that $(n-k)d > (T-t) > (n-k-1)d$ then $G_{f,(k,n)}(\mathbb{X}^n) \subset \triangle_{f,(k,n)}$, where $\triangle_{f,(k,n)} = \text{Conv}(G_{f,(k,n)}(V))$.

Sketch of Proof. To prove A) we first use the affine independence of $G_f(V)$ from Lemma 5.2 A) to construct $\triangle_f = \text{Conv}(G_f(V))$. Next we consider two subsets of designs in $\mathbb{X}^n$. Those for which $x_1 < t$ and those for which $x_1 > t$. When $x_1 < t$, we note that G_f will always map to the hyperplane $\pi_0 = 0$, which is coincident with a boundary of $\triangle_f$. When $x_1 > t$, the mapping G_f is injective and all its partial derivatives form linearly independent vectors. Therefore a proof similar to that for Theorem 5.1 can be used to show that, for log-concave f, we have $G_f(\mathbb{X}^n) \subset \triangle_f$.

The proof of B) is almost the same. By Lemma 5.2 B) we know that the simplex $\triangle_{f,(k,n)}$ exists. Since the design points x_1 to x_k can never appear in the support of f, we consider $G_{f,(k,n)}$. We again have two subsets of designs. The first has $x_{k+1} < t$ and the second $x_{k+1} > t$. When $x_{k+1} < t$, we note that $G_{f,(k,n)}$ maps to $\pi_k = 0$, which is a boundary of $\triangle_{f,(k,n)}$. When $x_{k+1} > t$, the mapping $G_{f,(k,n)}$ is injective with linearly independent partial derivative vectors. It can then be shown that, for log-concave f, we have $G_{f,(k,n)}(\mathbb{X}^n) \subset \triangle_{f,(k,n)}$. □

5.3.2 Type 2 Prior

The Type 2 prior density is the symmetric analogue to the Type 1 prior density. As a result, we simply state the equivalents to Lemma 5.2 and Theorem 5.2.

Lemma 5.3. *For a Type 2 prior density f:*
A) If the support $[0, t]$ is such that $t > (n-1)d$ then the set $G_f(V)$ is affinely independent.
B) If support $[0, t]$ is such that $(n-k)d > t > (n-k-1)d$ then the set $G_{f,(k,n)}(\{u_k, \ldots, u_n\})$ is affinely independent.

Theorem 5.3. *For a log-concave Type 2 prior density f:*
A) If the support $[0, t]$ is such that $t > (n-1)d$, then $G_f(\mathbb{X}^n) \subset \triangle_f$, where $\triangle_f = \mathrm{Conv}(G_f(V))$.
B) If the support $[0, t]$ is such that $(n-k)d > t > (n-k-1)d$, then $G_{f,(0,n-k)}(\mathbb{X}^n) \subset \triangle_{f,(0,n-k)}$, where $\triangle_{f,(0,n-k)} = \mathrm{Conv}(G_{f,(0,n-k)}(V))$.

5.3.3 Type 3 Prior

Lemma 5.4. *For a Type 3 prior density f:*
A) If the support $[t_0, t_T]$ is such that $t_T > (n-1)d$ and $(T - t_0) > (n-1)d$ then $G_f(V)$ is affinely independent.
B) If the support $[t_0, t_T]$ is such that $(n-l-1)d < t_T < (n-l)d$ and $(n-k-1)d < (T - t_0) < (n-k)d$ then $G_{f,(k,n-l)}(\{u_{n-k}, \ldots, u_l\})$ is affinely independent.

Sketch of proof. Using a proof similar to that of Lemma 5.1 we can show that $G_f(V)$, under the conditions in A), is affinely independent. In B), if either $t_T < (n-1)d$ or $(T - t_0) < (n-1)d$ then the set $G_f(V)$ is not affinely independent. We consider the case where there are k design points before t_0 and there are l design points after t_T. These two situations correspond to $(n-k-1)d < (T-t_0) < (n-k)d$ and $(n-l-1)d < t_T < (n-l)d$, respectively. Applying the method of Lemma 5.1 gives that $G_{f,(k,n-l)}(\{u_{n-k}, \ldots, u_l\})$ is affinely independent. □

Theorem 5.4. *For a Type 3 log-concave prior density f:*

A) If the support $[t_0, t_T]$ is such that $t_T > (n-1)d$ and $(T - t_0) > (n-1)d$, then $G_f(\mathbb{X}^n) \subset \triangle_f$, where $\triangle_f = \mathrm{Conv}(G_f(V))$. Also, if the support of f, $[t_0, t_T]$, is such that $(t_T - t_0) < (n-1)d$, then $G_f(\mathbb{X}^n) \subset \triangle_f$ regardless of whether or not f is log-concave.

B) If the support $[t_0, t_T]$ is such that $(n-l-1)d < t_T < (n-l)d$ and $(n-k-1)d < (T-t_0) < (n-k)d$, then $G_{f,(k,n-l)}(\mathbb{X}^n) \subset \triangle_{f,(k,n-l)}$ where $\triangle_{f,(k,n-l)} = \mathrm{Conv}(G_{f,(k,n-l)}(V))$. Furthermore, if $(t_T - t_0) < (n-k-l-1)d$ then $G_{f,(k,n-l)}(\mathbb{X}^n) \subset \triangle_{f,(k,n-l)}$, regardless of the shape of f.

Sketch of Proof. In A), for designs where either $x_1 < t_0$ and/or $x_n > t_T$ we map to $\pi_0 = 0$ and/or $\pi_n = 0$ which are facets of $\triangle_f$. If $(t_T - t_0) > (n-1)d$ then all the design points can fit in the support of f and, hence, G_f can map to places other than $\pi_0 = 0$ and $\pi_n = 0$. Since f is greater than zero on its support, the proof of Theorem 5.1 can be used for designs with $x_1 > t_0$ and $x_n < t_T$ to show that for, log-concave f, we have $G_f(\mathbb{X}^n) \subset \triangle_f$.

If $(t_T - t_0) < (n-1)d$ then either $x_1 < t_0$ or $x_n > t_T$ and we *always map* to $\pi_0 = 0$ or $\pi_n = 0$. Thus it is *not* necessary for f to be log-concave to ensure that $G_f(\mathbb{X}^n) \subset \triangle_f$.

In B), for designs where $x_{k+1} < t_0$ and/or $x_{n-l} > t_T$, we see that $G_{f,(k,n-l)}$ maps to $\pi_k = 0$ and/or $\pi_{n-l} = 0$. When the support $[t_0, t_T]$ is such that $(t_T - t_0) >$

$(n-k-l-1)d$ there exist designs with $x_{k+1} > t_0$ and $x_{n-1} < t_T$. Then we can prove that if f is log-concave then $G_{f,(k,n-l)}(\mathbb{X}^n) \subset \triangle_{f,(k,n-l)}$.

If the prior density has support $[t_0, t_T]$ such that $(t_T - t_0) < (n-k-l-1)d$ then $x_{k+1} < t_0$ and/or $x_{n-l} > t_T$. Therefore $G_{f,(k,n-l)}$ always maps to either $\pi_k = 0$ or $\pi_{n-l} = 0$, or both. Since $G_{f,(k,n-l)}$ maps to facets of $\triangle_{f,(k,n-l)}$, we have $G_{f,(k,n-l)}(\mathbb{X}^n) \subset \triangle_{f,(k,n-l)}$. Here there is *no* restriction on the shape of f. □

5.4 When f has Positive Probability of no Changepoint

In certain situations it is reasonable to assume that there might be no changepoint. Returning to our motivating example, it might happen that the blood pressure lowering treatment simply did not work. Anticipating this, we would like to allow for the possibility of no changepoint in our prior distribution for τ.

Here, we examine the situation when the prior density is continuous and differentiable on $[0, T)$ and has mass p_T at T. Recall that the event $\{\tau = T\}$ is equivalent to the event of no change. We consider Type 1, 2, and 3 prior densities as well as prior densities whose support coincides with $[0, T)$, except for each one we include a point mass at T.

Theorem 5.5. *When f is log-concave with mass p_T at T, Theorems 5.1, 5.2, 5.3 and 5.4 hold (the only exception being that by support we exclude the discrete mass at T).*

Sketch of Proof. First we look at the prior density everywhere greater than zero. Upon reflection, we see that our earlier proofs for Lemma 5.1 and Theorems 5.1 hold except that the facet $\pi_n = p_T$ now appears instead of $\pi_n = 0$. Following the same steps as before, we can show, for log-concave f, that $G_f(\mathbb{X}^n) \subset \triangle_f$.

Likewise, for the Type 1, 2, and 3 prior densities, the proofs remain the same except we have $\pi_n = p_T$ instead of $\pi_n = 0$. Note for Type 2 and 3 prior densities, when some design points always remain to the right of the support of the prior density, the dimension of the problem is reduced in exactly the same way. □

5.5 Optimal Designs for Design Criterion Functions Concave in π

In summary, for design criterion functions concave in π, combining Theorems 5.1, 5.2, 5.3, 5.4, and 5.5 we obtain Theorem 5.6.

Theorem 5.6. *If f is positive everywhere on $[0, T]$, or a Type 1, 2, or 3 log-concave prior density, with or without mass at T, then the optimal design for a design criterion function which is concave in π will be one of the designs in the set V.*

Remark 5.1. *For Type 1, 2, and 3 log-concave prior densities, under certain conditions on the support of f, we only need to consider a subset of V. Furthermore, in certain cases for a Type 3 prior density, the log-concavity is not necessary. These details are discussed further in Appendix A.*

Sketch of Proof. Theorems 5.1, 5.2, 5.3, 5.4, and 5.5 show us that the image of $\mathbb{X}^n$ under G_f or under some reduced dimensional form $G_{f,(\cdot,\cdot)}$ is a subset of the simplex $\triangle_f = \text{Conv}(G_f(V))$ or the simplex $\triangle_{f,(\cdot,\cdot)} = \text{Conv}(G_{f,(\cdot,\cdot)}(V))$, respectively. Obviously, it is the designs in V that map to the vertices of the simplices $\triangle_f$ and $\triangle_{f,(\cdot,\cdot)}$. From Corollary 3.1, we know one of the designs in V will be the design minimizing the concave design criterion function. □

5.6 Restriction to a Subset of Designs

When we test for a change or for a change in a particular interval, it will be necesssary to fix one or more design points. That is, we must restrict ourselves to a subset of the designs $\mathbb{X}^n$.

5.6.1 Fixing x_n at Time T

When we wish to test for a change it is necessary to fix x_n at T. Obviously, for such a testing problem it is essential that our prior density has mass p_T at T. With x_n at T, π_n will always equal p_T. We are then only concerned with the positions of the design points x_1 to x_{n-1}. All our previous work holds, except instead of starting with V and G_f we start with the set $\{u_1, \dots, u_n\}$ and the mapping $G_{f,(0,n-1)}$. We state the theorem below for priors having mass at T.

Theorem 5.7. *Consider designs with x_n at T, and suppose that f is a log-concave prior density with mass at T. If f is either of Type 1, 2, or 3 or is positive everywhere on $[0, T]$ then the optimal design for a design criterion function, which is concave in π, will be one of the designs in the set $\{u_1, \dots, u_n\}$.*

Remarks similar to the ones after Theorem 5.6 apply.

5.6.2 Fixing x_q and x_{q+1} at Times t_1 and t_2

As we will see, when we test for a change in the subinterval $[t_1, t_2]$, it will be necessary to fix two designs points. The interval $[t_1, t_2]$ is formed by fixing the positions of two adjacent design points, say, points x_q and x_{q+1}, such that x_q is at t_1 and x_{q+1} is at t_2. Now we are left with a subset of $\mathbb{X}^n$, where the points $(x_1, \dots, x_{q-1})$ occupy positions a distance d apart in the interval $[0, t_1 - d]$ and the points $(x_{q+2}, \dots, x_n)$ occupy positions a distance d apart in the interval $[t_2 + d, T]$.

Before presenting our next result, we introduce some more notation. Take a set C of q vectors in $(q-1)$-dimensional space: $(0, d, \dots, (q-2)d)$, $(0, d, \dots, (q-3)d, t_1-d)$, through to $(t_1-(q-1)d, \dots, t_1-d)$. We call these vectors $c_0, c_1, \dots, c_{q-1}$, respectively. Thus, C is defined as

$$C = \{c_0, c_1, \dots, c_{q-1}\}. \tag{5.14}$$

Similarly, let D be a set $(n-q)$ vectors in $(n-q-1)$-dimensional space: $(t_2+d, \dots, t_2+(n-q-1)d)$, $(t_2+d, \dots, t_2+(n-q-2)d, T)$, through to $(T-(n-q-2)d, \dots, T-d, T)$. We call these vectors $d_0, d_1, \dots, d_{n-q-1}$, respectively. Thus, D is defined as

$$D = \{d_0, d_1, \dots, d_{n-q-1}\}. \tag{5.15}$$

Theorem 5.8. *Suppose that x_q and x_{q+1} are fixed at times t_1 and t_2, respectively, and f is a log-concave prior density with mass at T. Let f be either of Type 1, 2, or 3 or positive everywhere on $[0,T]$. Then a design criterion function, which is concave in π, has an optimal design which is a Cartesian product of an element of C and an element of D. The sets C and D are defined in equations (5.14) and (5.15) respectively.*

Sketch of Proof. Denote the subsets $\mathbb{X}^n_{(1,q-1)}$ and $\mathbb{X}^n_{(q+2,n)}$ as the designs in $\mathbb{X}^n$ restricted to the coordinates $(x_1, \dots, x_{q-1})$ and $(x_{q+2}, \dots, x_n)$, respectively. If the prior density is everywhere greater than zero on $[0,T]$ we can show that the sets $G_{f,(0,q-1)}(C)$ and $G_{f,(q+1,n)}(D)$ are affinely independent. Hence we have the simplices $\triangle_{f,(0,q-1)} = \text{Conv}(G_{f,(0,q-1)}(C))$ and $\triangle_{f,(q+1,n)} = \text{Conv}(G_{f,(q+1,n)}(D))$. As with Theorem 5.1, we can prove that if f is log-concave then $G_{f,(0,q-1)}(\mathbb{X}^n_{(0,q-1)}) \subset \triangle_{f,(0,q-1)}$ and $G_{f,(q+1,n)}(\mathbb{X}^n_{(q+1,n)}) \subset \triangle_{f,(q+1,n)}$. Therefore, for log-concave prior densities, the points $\pi = (\pi_0, \dots, \pi_{q-1}, \hat{\pi}_q, \pi_{q+1}, \dots, \pi_n)$ lie in the Cartesian product of $G_{f,(0,q-1)}(\mathbb{X}^n_{(0,q-1)})$ and $G_{f,(q+1,n)}(\mathbb{X}^n_{(q+1,n)})$, both of which are subsets of a simplex and containing the vertices of the simplex. By Corollary 3.2, we have our result.

The above also holds if the prior density has mass at T. Similar proofs follow for Type 1, 2, and 3 prior densities both with and without mass at T. In the case of a Type 1, 2, or 3 prior density one has to consider situations where the dimension of the problem is reduced. □

Note that, according to Theorem 5.8, the optimal design for a concave design criterion function in π when testing for a change in the subinterval $[t_1, t_2]$ will be one of the designs placing design points as far as possible towards t_1 and t_2. Additional design points may be placed towards 0 and/or towards T.

Chapter 6

Optimal Designs for the Single-Path Changepoint Problem

Here we find optimal designs for the single-path changepoint problem introduced in Section 4.1. The designs we obtain are optimal for testing for a change, testing for a change in a sub-interval, and estimating the before-and-after-change means. Furthermore, by taking a convex combination, we can combine criterion functions to find designs that are optimal for both testing for a change and estimating the before-and-after-change means.

Much of the necessary ground work has been done in Chapter 5, where we considered the shape of $G_f(\mathbb{X}^n)$. The results in this chapter follow immediately from the concluding theorems of Chapter 5. Theorem 5.7 leads to optimal design results to test for a change, Theorem 5.8 to test for a change in a subinterval, and Theorem 5.6 to estimate the before-and-after-change means.

To use these theorems we simply need to prove that the design criterion functions, discussed in Chapter 2 are concave functions of the design measure π. This is the goal of the present chapter.

As presented in Chapter 2 our design criterion functions for the testing problems are the commonly used Bayes risk based on generalized 0-1 loss and the Spezzaferri criterion function for model discrimination. Our design criterion function for estimating the before-and-after-change means is the well-known Bayes risk based on squared error loss. These are the three criterion functions considered in this thesis. However, by Theorems 5.7, 5.8, and 5.6, our optimal design results apply to *any* design criterion function which is concave in π.

We begin in Section 6.1 by presenting the single-path model in a more general format than was presented in Section 4.1. We then calculate the posterior distributions of the model emphasizing their dependence on π. In Section 6.2 we consider designs that are optimal for testing if there has been a changepoint and for testing if the changepoint occurred in a specific interval. Section 6.3 concerns estimation. We discuss why it is difficult to find the designs that are optimal for estimating the changepoint. Next we consider designs that are optimal for estimating the before-and-after-change means. Finally, in Section 6.4, we combine criterion functions for testing and estimation.

6.1 Model and Dependence on π

Ultimately, we wish to consider our design criterion functions as functions of the design measure π. Hence, we begin here by investigating how the likelihood and posterior densities depend on π. Again, we do not distinguish between random variables and their realized values. Whether a quantity is a random variable or not will be evident from the context.

The likelihood, which was first stated in equation (4.1) in its conditionally independent form, is repeated here in a more general form, allowing for a correlation between observations. It is assumed (perhaps restrictively) that the correlation structure does not depend on the design. Since we do not allow the densities to depend on the design x, the likelihood will have no further dependence on x or τ except through τ_x. Recall that the event $\{\tau_x = k\}$ represents the number of observations taken up to and including the changepoint. We will split the observation vector y into two vectors y^1 and y^2, where y^1 represents the first k observations from the μ_1 distribution $f(\cdot|\mu_1)$ and y^2 represents the last $n-k$ observations from the μ_2 distribution $f(\cdot|\mu_2)$. The joint density thus factors as

$$f(y|\mu_1,\mu_2,\tau_x = k) = f(y^1|\mu_1)f(y^2|\mu_2). \tag{6.1}$$

Our model also includes arbitrary and marginal prior densities $f(\mu_1,\mu_2)$ and $f(\tau)$. The random variables μ_1 and μ_2 are assumed to be independent of τ and hence the joint prior density is specified as $f(\mu_1,\mu_2)f(\tau)$.

We now consider the posterior densities of our single-path changepoint model. The conditional density of y given $\tau_x = k$ is

$$\begin{aligned} f(y|\tau_x = k) &= \int\int f(y,\mu_1,\mu_2|\tau_x = k)d\mu_1 d\mu_2 \\ &= \int\int f(y|\mu_1,\mu_2,\tau_x = k)f(\mu_1,\mu_2)d\mu_1 d\mu_2. \end{aligned} \tag{6.2}$$

The density $f(y|\tau_x = k)$ is used to obtain the posterior densities $f(\mu_1|y,\tau_x = k)$ and $f(\mu_2|y,\tau_x = k)$ as follows:

$$\begin{aligned} f(\mu_1|y,\tau_x = k) &= \frac{\int f(y,\mu_1,\mu_2,\tau_x = k)d\mu_2}{f(y|\tau_x = k)\pi_k} \\ &= \frac{\int f(y,\mu_1,\mu_2|\tau_x = k)d\mu_2}{f(y|\tau_x = k)} \\ &= \frac{\int f(y|\mu_1,\mu_2,\tau_x = k)f(\mu_1,\mu_2)d\mu_2}{f(y|\tau_x = k)} \end{aligned} \tag{6.3}$$

and similarly

$$f(\mu_2|y,\tau_x = k) = \frac{\int f(y|\mu_1,\mu_2,\tau_x = k)f(\mu_2,\mu_1)d\mu_1}{f(y|\tau_x = k)}. \tag{6.4}$$

The marginal distribution of y is a mixture distribution with weights equal to the components of π:

$$f(y) = \sum_{k=0}^{n} f(y|\tau_x = k)\pi_k. \tag{6.5}$$

Next, we consider the density $f(\tau_x = k|y)$, which will obviously depend on π.

$$f(\tau_x = k|y) = \frac{f(y|\tau_x = k)\pi_k}{\sum_{l=0}^{n} f(y|\tau_x = l)\pi_l} \tag{6.6}$$

In the next sections we consider how our design criterion functions depend on π and prove that they are concave functions of π.

6.2 Optimal Designs for Testing

Here, we consider designs that are optimal for testing for a change both generally and in a specific subinterval $[t_1, t_2]$. We use both the Bayes risk based on generalized 0-1 loss and the Spezzaferri condition. As we saw in Chapter 5, for these testing problems it is necessary to fix one or two design points, thereby restricting ourselves to a subset of the design space $\mathbb{X}^n$.

6.2.1 Optimal Design for Testing for a Change

In this section we find the optimal design for choosing between the models $\{\tau = T\}$ (no change occurs) versus $\{\tau < T\}$ (a change occurs). Our prior distribution for τ is of the type discussed in Section 5.4, where we have a point mass p_T at T to allow for the possibility of no change. Referring to our motivational Example 1.1 in Chapter 1, we would conduct such a test if we were interested in knowing whether or not the blood pressure treatment has an effect.

Since we can only take measurements at n locations in the continuous interval $[0, T]$, we have to insist that our last measurement x_n be taken at T in order to make inference about possible events at that point. By fixing x_n at the location T, we are fixing the value of π_n at p_T and thereby reducing the problem by one dimension. Consequently, we will search for the optimal design for a test of change from amongst the set of all possible positions of the design points x_1 through to x_{n-1} in the interval $[0, T-d]$. In other words, we are considering the subset of $\mathbb{X}^n$ where $x_n = T$ for every design in the subset.

Now, we prove that the Bayes risk based on generalized 0-1 loss and the Spezzaferri criterion functions are concave over the set of vectors

$$\{(\pi_0, \ldots, \pi_{n-1}) \mid \sum_{k=1}^{n-1} \pi_k = 1 - p_T\}.$$

Let $M_{\tau<T}$ denote the model corresponding to the event $\{\tau < T\}$ and $M_{\tau=T}$ denote the model corresponding to the event $\{\tau = T\}$. Obviously, with x_n fixed at T, the events $\{\tau < T\}$ and $\{\tau_x < n\}$ are equivalent and the events $\{\tau = T\}$ and $\{\tau_x = n\}$ are equivalent. Hence we have,

$$f(y|M_{\tau<T}) = f(y|\tau_x < n) = \frac{\sum_{k=0}^{n-1} f(y|\tau_x = k)\pi_k}{\sum_{l=0}^{n-1} \pi_l} \tag{6.7}$$

and

$$f(y|M_{\tau=T}) = f(y|\tau_x = n). \tag{6.8}$$

From expression (2.3) of Section 2.2, the Bayes risk based on generalized 0-1 loss for testing the model $M_{\tau=T}$ versus $M_{\tau<T}$ with a null hypothesis of no change is

$$K_0 \int_{R_0} P(M_{\tau<T}) f(y|M_{\tau<T}) dy + K_1 \int_{R_1} P(M_{\tau=T}) f(y|M_{T=\tau}) dy. \tag{6.9}$$

Lemma 6.1. *With x_n fixed at T, the Bayes risk based on generalized 0-1 loss for testing for a changepoint in the single-path model is a linear function of the vector $(\pi_0, \ldots, \pi_{n-1})$.*

Proof. Substituting for $P(M_{\tau=T})$, $P(M_{\tau<T})$, $f(y|M_{T=\tau})$ and $f(y|M_{\tau<T})$, we re-write the Bayes risk (6.9) as

$$K_0 \int_{R_0} \sum_{k=1}^{n-1} \pi_k f(y|\tau_x = k)dy + K_1 \int_{R_1} p_T f(y|\tau_x = n)dy. \tag{6.10}$$

We see that the Bayes risk (6.10) is linear in $(\pi_0, \ldots, \pi_{n-1})$. □

From expression (2.8) of Section 2.4, the Spezzaferri criterion function reduces to minimizing

$$\int \frac{f(y|M_{\tau=T})f(y|M_{\tau<T})}{f(y)}dy. \tag{6.11}$$

Lemma 6.2. *With x_n fixed at T, the Spezzaferri criterion function for testing for a change in the single-path model is a concave function of the vector $(\pi_0, \ldots, \pi_{n-1})$.*

Proof. Substituting expressions (6.7) and (6.8) into (6.11), the Spezzaferri criterion function becomes

$$\int \frac{f(y|\tau_x = n)(\sum_{k=0}^{n-1} f(y|\tau_x = k)\pi_k)}{(\sum_{l=0}^{n-1} \pi_l)(\sum_{r=0}^{n} f(y|\tau_x = r)\pi_r)}dy. \tag{6.12}$$

Using Theorem 3.6, we show the concavity of integral (6.12) by showing that its integrand is concave for all y. We drop the factor $f(y|\tau_x = n)$ and the constant $\sum_{l=0}^{n-1} \pi_l = 1 - p_T$ since they do not affect concavity.

Denote the integrand by g. According to Lemma 3.3, to prove concavity of (6.11), we need to show

$$g(t_a\pi_{(a)} + t_b\pi_{(b)}) \geq t_a g(\pi_{(a)}) + t_b g(\pi_{(b)}), \text{ for } t_a + t_b = 1 \tag{6.13}$$

for any two vectors $\pi_{(a)}$ and $\pi_{(b)}$. Letting $\pi_{(a),k}$ and $\pi_{(b),k}$ be the kth components of $\pi_{(a)}$ and $\pi_{(b)}$ respectively, we simplify notation by letting $A = \sum_{k=0}^{n-1} f(y|\tau_x = k)\pi_{(a),k}$, and $B = \sum_{k=0}^{n-1} f(y|\tau_x = k)\pi_{(b),k}$. We must then show that

$$\frac{t_aA + t_bB}{t_aA + t_bB + f(y|\tau_x = n)\pi_n} \geq \frac{t_aA}{A + f(y|\tau_x = n)\pi_n} + \frac{t_bB}{B + f(y|\tau_x = n)\pi_n} \tag{6.14}$$

which is equivalent to showing

$$t_a\left(\frac{A}{t_aA+t_bB+f(y|\tau_x=n)p_T}-\frac{A}{A+f(y|\tau_x=n)p_T}\right)+$$
$$t_b\left(\frac{B}{t_aA+t_bB+f(y|\tau_x=n)p_T}-\frac{B}{B+f(y|\tau_x=n)p_T}\right)\geq 0.$$

Further simplifying our notation we set $C=f(y|\tau_x=n)p_T$. We need to prove that,

$$\frac{t_at_b}{t_aA+t_bB+C}\left(\frac{A(A-B)}{A+C}+\frac{B(B-A)}{B+C}\right)\geq 0$$

which, in turn, becomes

$$\frac{t_at_b(A-B)^2}{t_aA+t_bB+C}\left(\frac{C}{(A+C)(B+C)}\right)\geq 0.$$

The result follows as all the quantities on the left-hand side of the above inequality are positive. □

Recall the designs u_i, $i=1,\ldots,n$, defined in Section 4.2. These are used in the following theorem.

Theorem 6.1. *Consider a single-path changepoint problem with a log-concave prior distribution for the changepoint τ and the design point x_n fixed at T. With respect to the Bayes risk based on generalized 0-1 loss and the Spezzaferri criterion function for model discrimination, the optimal design for testing for a change is one of the designs in the set $\{u_1,\ldots,u_n\}$.*

Proof. By Lemmas 6.1 and 6.2, the generalized 0-1 Bayes risk and the Spezzaferri criterion functions are concave. The result follows directly from Theorem 5.7. □

6.2.2 Optimal Design for Testing for a Change in a Subinterval

Recollecting our motivational example in Chapter 1, we might be interested in testing if the treatment has an effect during an interval shortly after it is administered. For instance, if the treatment was to be administered at time t_1, we might want to assess if a change in mean blood pressure occured in the interval $[t_1, t_2]$ where t_2 is greater than t_1. Here we present optimal designs for such a test.

The interval $[t_1, t_2]$ is constructed by fixing the positions of two adjacent design points x_q and x_{q+1}, such that x_q is at t_1 and x_{q+1} is at t_2. Such intervals are discussed in Section 5.6.2.

We consider optimal designs for testing $M_{\tau\in[t_1,t_2]^C}$ versus $M_{\tau\in[t_1,t_2]}$. Again we use the Bayes risk based on generalized 0-1 loss and the Spezzaferri criterion function. With x_q and x_{q+1} fixed at t_1 and t_2, respectively, our interest lies in

$$f(y|M_{\tau\in[x_q,x_{q+1}]^C}) = f(y|\tau_x \neq q) = \frac{\sum_{k\neq q} f(y|\tau_x = k)\pi_k}{\sum_{l\neq q} \pi_l} \tag{6.15}$$

and

$$f(y|M_{\tau\in[x_q,x_{q+1}]}) = f(y|\tau_x = q). \tag{6.16}$$

Due to the fixed positions of x_q and x_{q+1}, the value of π_q is $\int_{t_1}^{t_2} f$, which we will denote as p_{t_1,t_2}.

If the null hypothesis is that τ occurs in the interval $[t_1, t_2]$, the Bayes risk for this problem is

$$K_0 \int_{R_0} P(M_{\tau\in[t_1,t_2]^C}) f(y|M_{\tau\in[t_1,t_2]^C}) dy + K_1 \int_{R_1} P(M_{\tau\in[t_1,t_2]}) f(y|M_{\tau\in[t_1,t_2]}) dy. \tag{6.17}$$

Lemma 6.3. *Consider a single-path changepoint problem with x_q and x_{q+1} fixed at t_1 and t_2 respectively. The Bayes risk based on generalized 0-1 loss for testing for a change in $[t_1, t_2]$ is a linear function of the vector $(\pi_0, \ldots, \pi_{q-1}, \hat{\pi}_q, \pi_{q+1}, \ldots, \pi_n)$.*

Proof. Substituting Expressions (6.15) and (6.16) into the Bayes risk (6.17) we have,

$$K_0 \int_{R_0} \sum_{k\neq q} \pi_k f(y|\tau_x = k) dy + K_1 \int_{R_1} p_{t_1,t_2} f(y|\tau_x = q) dy \tag{6.18}$$

which is obviously a linear function of the vector $(\pi_0, \ldots, \pi_{q-1}, \hat{\pi}_q, \pi_{q+1}, \ldots, \pi_n)$. □

Next, we consider the Spezzaferri criterion function (2.8) which reduces to minimizing

$$\int \frac{f(y|M_{\tau\in[t_1,t_2]})f(y|M_{\tau\in[t_1,t_2]^C})}{f(y)}dy. \tag{6.19}$$

Lemma 6.4. *Consider the single-path changepoint problem with x_q and x_{q+1} fixed at t_1 and t_2 respectively. The Spezzaferri criterion function for testing for a change in $[t_1, t_2]$ is a concave function of the vector $(\pi_0, \ldots, \pi_{q-1}, \hat{\pi}_q, \pi_{q+1}, \ldots, \pi_n)$.*

Proof. Using (6.15) and (6.16) again, the Spezzaferri condition becomes

$$\int \frac{f(y|\tau_x = q)(\sum_{k\neq q} f(y|\tau_x = k)\pi_k)}{(\sum_{l\neq q}\pi_l)(\sum_{r=0}^{n} f(y|\tau_x = r)\pi_r)}dy. \tag{6.20}$$

This criterion is formally equivalent to (6.12) with index n replaced by index q. The rest of the proof follows through as in the proof of Lemma 6.2. □

Recall C and D, defined in (5.14) and (5.15), respectively.

Theorem 6.2. *Consider a single-path changpoint problem with a log-concave prior for τ and x_q and x_{q+1} fixed at t_1 and t_2 respectively. With respect to the generalized 0-1 Bayes risk and the Spezzaferri criterion function, the optimal design is one of the designs which is a Cartesian product of an element of C and an element of D.*

Proof. This result follows directly from Theorem 5.8, which states that the optimal design for a design criterion function concave in π is a Cartesian product of an element from C and an element from D, and Lemmas 6.3 and 6.4, which show the generalized 0-1 Bayes risk and the Spezzaferri criterion function are concave functions of π. □

Now in the above discussion we have required that a design point be placed at each of the points t_1 and t_2. We have not however, given any guidance as to which points should be placed at t_1 and t_2. To find the "optimal" q we suggest forming the interval $[t_1, t_2]$ with x_q and x_{q+1} for each q from 1 to $n-1$ and finding the "optimal design" for the chosen criterion function. Then among the $n-1$ optimal designs, select the one that is associated to the lowest value of the design criterion function of interest.

6.3 Designs and Optimal Designs for Estimation

In this section, we find optimal designs for estimating the before-and-after-change means. We conclude by discussing the optimal design problem for estimating the changepoint location.

6.3.1 Optimal Designs for Estimating the Before-and-After-Change Means

Next we consider optimal designs for estimating the before-and-after-change means. The design criterion function we use is the Bayes risk based on squared error loss, introduced in Section 2.2. When finding the optimal design for estimating the before-and-after-change means, we would usually assume a changepoint has occurred. Hence, we would have a prior on τ with no mass at T.

Recall from expression (2.2), that the Bayes risk based on squared error loss is the posterior variance of the parameter of interest, averaged over the anticipated data. Since we have two means to estimate, we use the sum

$$R = \int Var(\mu_1|y)f(y)dy + \int Var(\mu_2|y)f(y)dy, \tag{6.21}$$

to define the risk. Following Zhou (1997) we use the well-known identity (6.22) to divide the Bayes risk (6.21) into four terms.

$$Var(\mu|y) = E_{\tau_x=k|y}(Var(\mu|y,\tau_x=k)) + Var_{\tau_x=k|y}(E(\mu|y,\tau_x=k)) \tag{6.22}$$

The Bayes risk which we denote as R, re-expressed as the sum of the four terms, is

$$\begin{aligned} R = &\int E_{\tau_x=k|y}(Var(\mu_1|y,\tau_x=k))f(y)dy \\ &+ \int Var_{\tau_x=k|y}(E(\mu_1|y,\tau_x=k))f(y)dy \\ &+ \int E_{\tau_x=k|y}(Var(\mu_2|y,\tau_x=k))f(y)dy \\ &+ \int Var_{\tau_x=k|y}(E(\mu_2|y,\tau_x=k))f(y)dy. \end{aligned} \tag{6.23}$$

Denote these four integrals by R_1, R_2, R_3, and R_4, respectively. As the terms R_1 and R_3 have the same structure, and the terms R_2 and R_4 have the same structure, we first consider R_1 and R_3 together, and then consider R_2 and R_4 together. In particular, we ascertain how these terms depend on π. Note that the terms R_1 and R_3 describe the within-model variability, while the terms R_2 and R_4 describe the between-model variability. The within-model variability refers to the variability around the means μ_1 and μ_2 given a fixed changepoint, and the between-model variability refers to the extra variability induced by the uncertainty of the location of the changepoint.

Starting with R_1 and R_3, we prove the crucial result that the Bayes risk based on squared error loss is a concave function of π. By Fubini's theorem, R_1 and R_3 can be re-expressed as

$$R_1 = \sum_{k=0}^{n} E_{y|\tau_x=k}(Var(\mu_1|y,\tau_x=k))\pi_k \tag{6.24}$$

and

$$R_3 = \sum_{k=0}^{n} E_{y|\tau_x=k}(Var(\mu_2|y,\tau_x=k))\pi_k. \tag{6.25}$$

It follows that $R_1 + R_3$ is equivalent to the generalized Lauter's criterion function introduced in Zhou et al. (2003), where a Bayes risk based on squared error loss is used for each value of τ_x to estimate μ_1 and μ_2. This observation has been made before in Zhou (1997).

Lemma 6.5. *Suppose we have the Bayes risk based on squared error loss in the single-path changepoint problem for estimating the before-and-after-change means. Then the terms R_1 and R_3 of the Bayes risk are linear functions of the coordinates of π.*

Proof. We consider only R_1 since R_3 is dealt with in exactly the same way. First, recall that the density $f(\mu_1|y, \tau_x = k)$ has no dependence on π (see (6.3)). Therefore, the expectation $E(\mu_1|y, \tau_x = k) = \int \mu_1 f(\mu_1|y, \tau_x = k)d\mu_1$ has no dependence on π, and hence the variance $Var(\mu_1|y, \tau_x = k) = \int(\mu_1 - E(\mu_1|y, \tau_x = k))^2 f(\mu_1|y, \tau_x = k)d\mu_1$ also has no dependence on π. Observing that the density, $f(y|\tau_x = k)$, in (6.2), does not depend on π, we find that $E_{y|\tau_x=k}(Var(\mu_1|y, \tau_x = k))$ does not depend on π. □

Knowing that R_1 and R_3 are simply linear combinations of the components of π, we find they are concave. Intuitively, we expect the minimum of $E_{y|\tau_x=k}(Var(\mu_1|y, \tau_x = k))$ to occur when $\tau_x = n$, that is, when all the measurements are taken before the change. Hence, if we minimize over the set of all possible designs, including designs which allow design points to crowd together, we expect the minimum of R_1 to occur at the vertex of S^n with $\pi_n = 1$. This is the vertex of S^n, corresponding to all observations at the endpoint 0. Using the same reasoning, we would expect the minimum of R_3 to occur when all the points are at T, that is, at the vertex with $\pi_0 = 1$. As we will see in our numerical simulations in Chapter 7, this is exactly what happens.

Next we consider R_2 and R_4. We begin by rewriting R_2 and R_4 as

$$R_2 = \int \sum_{k=0}^{n} \left(E(\mu_1|y, \tau_x = k) - \sum_{l=0}^{n} E(\mu_1|y, \tau_x = l) f(\tau_x = l|y) \right)^2 \times f(\tau_x = k|y) f(y) dy \tag{6.26}$$

and

$$R_4 = \int \sum_{k=0}^{n} \left(E(\mu_2|y, \tau_x = k) - \sum_{l=0}^{n} E(\mu_2|y, \tau_x = l) f(\tau_x = l|y) \right)^2 \times f(\tau_x = k|y) f(y) dy. \tag{6.27}$$

From equations (6.5) and (6.6) we see that $f(y)$ and $f(\tau_x = k|y)$ are functions of π so that R_2 and R_4 are also functions of π. Recalling Theorem 3.6, if we prove that the integrands for R_2 and R_4 are concave in π for any value of y then the integrals R_2 and R_4 will also be concave in π. Therefore, we begin by proving the integrands of R_2 and R_4 are concave in π. Our first step is to introduce a simpler form for the integrands of R_2 and R_4 in Lemma 6.6.

Lemma 6.6. *The integrands of R_2 and R_4 also have the form*

$$\sum_{k=0}^{n}\sum_{l=0}^{k-1}(E(\mu_1|y,\tau_x = k) - E(\mu_1|y,\tau_x = l))^2\,\frac{f(y|\tau_x = l)\pi_l f(y|\tau_x = k)\pi_k}{\sum_{r=0}^{n} f(y|\tau_x = r)\pi_r}, \tag{6.28}$$

for R_2, and

$$\sum_{k=0}^{n}\sum_{l=0}^{k-1}(E(\mu_2|y,\tau_x = k) - E(\mu_2|y,\tau_x = l))^2\,\frac{f(y|\tau_x = l)\pi_l f(y|\tau_x = k)\pi_k}{\sum_{r=0}^{n} f(y|\tau_x = r)\pi_r}, \tag{6.29}$$

for R_4.

Proof. Since the proof for R_4 follows exactly the same steps as the proof for R_2 we present only the proof for R_2. To simplify notation we denote $g(k) = E(\mu_1|y,\tau_x = k)$. The integrand for R_2 is written as

$$\sum_{k=0}^{n}\left(g(k) - \sum_{l=0}^{n} g(l)\frac{f(y|\tau_x = l)\pi_l}{f(y)}\right)^2 \frac{f(y|\tau_x = k)\pi_k}{f(y)}f(y).$$

Factoring out $\left(\frac{1}{f(y)}\right)^2$ from the squared term leaves us with

$$\sum_{k=0}^{n}\left(g(k)f(y) - \sum_{l=0}^{n} g(l)f(y|\tau_x = l)\pi_l\right)^2 \frac{f(y|\tau_x = k)\pi_k}{f(y)^2}.$$

Substituting $\sum_{s=0}^{n} f(y|\tau_x = s)\pi_s$ for $f(y)$ and rearranging, we obtain

$$\sum_{k=0}^{n}\left(\sum_{l=0}^{n}(g(k) - g(l))f(y|\tau_x = l)\pi_l\right)^2 \frac{f(y|\tau_x = k)\pi_k}{(\sum_{s=0}^{n} f(y|\tau_x = s)\pi_s)^2}.$$

Expanding the squared summation we obtain,

$$\begin{aligned}
&\sum_{k=0}^{n}\left(\sum_{l=0}^{n}(g(k)-g(l))^2 f(y|\tau_x=l)^2\pi_l^2\right.\\
&\left.+\,2\sum_{l=0}^{n}\sum_{r=0}^{l-1}(g(k)-g(l))(g(k)-g(r))f(y|\tau_x=l)\pi_l f(y|\tau_x=r)\pi_r\right)\\
&\times\frac{f(y|\tau_x=k)\pi_k}{(\sum_{s=0}^{n} f(y|\tau_x=s)\pi_s)^2}.
\end{aligned}\tag{6.30}$$

Working with the cross terms we expand $2(g(k)-g(l))(g(k)-g(r))$, $2(g(l)-g(k))(g(l)-g(r))$, and $2(g(r)-g(l))(g(r)-g(k))$ and by factoring out $f(y|\tau_x=l)\pi_l f(y|\tau_x=r)\pi_r f(y|\tau_x=k)\pi_k$ we combine them into $2(g(k)^2+g(l)^2+g(r)^2-g(k)g(l)-g(k)g(r)-g(l)g(r))$, hence reducing the number of non-zero terms in the second summation from $\frac{(n+1)n(n-1)}{2}=(n-1)\binom{n+1}{2}$ to $\frac{(n+1)n(n-1)}{6}=\binom{n+1}{3}$. Letting $\sum_{(k,l)}$ indicate the sum over all $\binom{n+1}{2}$ pairs such that $k\neq l$ and $\sum_{(k,l,r)}$ indicate the sum over all the $\binom{n+1}{3}$ triplets such that $k>l>r$, expression (6.30) becomes

$$\begin{aligned}
&\sum_{(k,l)}(g(k)-g(l))^2\\
&\times\frac{f(y|\tau_x=l)^2\pi_l^2 f(y|\tau_x=k)\pi_k+f(y|\tau_x=l)\pi_l f(y|\tau_x=k)^2\pi_k^2}{(\sum_{s=0}^{n} f(y|\tau_x=s)\pi_s)^2}\\
&+\sum_{(k,l,r)}2(g(k)^2+g(l)^2+g(r)^2-g(k)g(l)-g(k)g(r)-g(l)g(r))\\
&\times\frac{f(y|\tau_x=l)\pi_l f(y|\tau_x=r)\pi_r f(y|\tau_x=k)\pi_k}{(\sum_{s=0}^{n} f(y|\tau_x=s)\pi_s)^2}.
\end{aligned}$$

Next, from each $2(g(k)^2+g(l)^2+g(r)^2-g(k)g(l)-g(k)g(r)-g(l)g(r))$ term we extract the terms $(g(k)-g(l))^2$, $(g(k)-g(r))^2$, and $(g(r)-g(l))^2$. By doing so we increase the number of terms in our second summation from $\binom{n+1}{3}$ back to the original number of terms, namely $3\binom{n+1}{3}=(n-1)\binom{n+1}{2}$. Collecting the $(n-1)$ terms $f(y|\tau_x=l)\pi_l f(y|\tau_x=r)\pi_r f(y|\tau_x=k)\pi_k$, and multiplying the terms $(g(k)-g(l))^2$, we are left with $\binom{n+1}{2} = \frac{3\binom{n+1}{3}}{(n-1)}$ terms in the second summation:

$$\left(\sum_{(k,l)} (g(k) - g(l))^2 (f(y|\tau_x = l)^2 \pi_l^2 f(y|\tau_x = k)\pi_k + f(y|\tau_x = l)\pi_l f(y|\tau_x = k)^2 \pi_k^2) \right.$$
$$\left. + \sum_{(k,l)} (g(k) - g(l))^2 \left(\sum_{r=0, r\neq k \neq l} f(y|\tau_x = l)\pi_l f(y|\tau_x = r)\pi_r f(y|\tau_x = k)\pi_k \right) \right)$$
$$\times \frac{1}{(\sum_{s=0}^n f(y|\tau_x = s)\pi_s)^2}$$

The $\binom{n+1}{2}$ terms in the first summation can now be combined with the $\binom{n+1}{2}$ terms in the second summation. Cancellation of $\sum_{s=0}^n f(y|\tau_x = s)\pi_s$ from the numerator and denominator, leads to the desired form

$$\sum_{(k,l)} (g(k) - g(l))^2 \frac{f(y|\tau_x = l)\pi_l f(y|\tau_x = k)\pi_k}{\sum_{r=0}^n f(y|\tau_x = r)\pi_r}. \tag{6.31}$$

Using exactly the same steps for R_4 and letting $h(k) = E(\mu_2|y, \tau_x = k)$, we find

$$\sum_{(k,l)} (h(k) - h(l))^2 \frac{f(y|\tau_x = l)\pi_l f(y|\tau_x = k)\pi_k}{\sum_{r=0}^n f(y|\tau_x = r)\pi_r}. \tag{6.32}$$

□

Before proving that R_2 and R_4 are concave we prove two further lemmas.

Lemma 6.7. *Let D be a $(n+1) \times (n+1)$ matrix with entries $D_{ij} = (d_i - d_j)^2$. Let $W = (w_0, \ldots, w_n)$ be such that $\sum_{i=0}^n w_i = 0$. Then $WDW' \leq 0$.*

Proof. We compute

$$\begin{aligned}
WDW' &= \sum_{i,j} w_i D_{i,j} w_j \\
&= \sum_{i,j} w_i (d_i - d_j)^2 w_j \\
&= \sum_{i,j} w_i d_i^2 w_j - 2\sum_{i,j} d_i w_i d_j w_j + \sum_{i,j} w_i d_j^2 w_j \\
&= -2(\sum_i d_i w_i)^2 \\
&\leq 0.
\end{aligned}$$

The proof is now complete. □

Lemma 6.8. *For all $d_i \in \mathbb{R}$ and all $x_i,\ y_i \in \mathbb{R}^+$, $i = 0, \dots, n$, we have*

$$\begin{aligned}
&\frac{\sum_{i=0}^n \sum_{j=0}^{i-1} (d_i - d_j)^2 (x_i + y_i)(x_j + y_j)}{\sum_{l=0}^n (x_l + y_l)} \\
&\geq \frac{\sum_{i=0}^n \sum_{j=0}^{i-1} (d_i - d_j)^2 x_i x_j}{\sum_{l=0}^n x_l} + \frac{\sum_{i=0}^n \sum_{j=0}^{i-1} (d_i - d_j)^2 y_i y_j}{\sum_{l=0}^n y_l}.
\end{aligned} \tag{6.33}$$

Proof. Expanding the terms in the numerator on the left-hand side of (6.33) and combing the two terms on the right-hand side, we find that the inequality (6.33) is equivalent to

$$\begin{aligned}
&\frac{\sum_{i=0}^n \sum_{j=0}^{i-1} (d_i - d_j)^2 x_i x_j + \sum_{i=0}^n \sum_{j=0}^{i-1} (d_i - d_j)^2 y_i x_j}{\sum_{l=0}^n (x_l + y_l)} \\
&+ \frac{\sum_{i=0}^n \sum_{j=0}^{i-1} (d_i - d_j)^2 x_i y_j + \sum_{i=0}^n \sum_{j=0}^{i-1} (d_i - d_j)^2 y_i y_j}{\sum_{l=0}^n (x_l + y_l)} \\
&\geq \frac{\sum_{i=0}^n \sum_{j=0}^{i-1} (d_i - d_j)^2 x_i x_j (\sum_{l=0}^n y_l) + \sum_{i=0}^n \sum_{j=0}^{i-1} (d_i - d_j)^2 y_i y_j (\sum_{l=0}^n x_l)}{(\sum_{l=0}^n x_l)(\sum_{l=0}^n y_l)}.
\end{aligned} \tag{6.34}$$

After cross-multiplication and cancellation of terms on each side, the inequality (6.34) becomes

$$\begin{aligned}&\left(\sum_{l=0}^{n} x_l\right)\left(\sum_{l=0}^{n} y_l\right)\left(\sum_{i=0}^{n}\sum_{j=0}^{i-1}(d_i-d_j)^2 y_i x_j+\sum_{i=0}^{n}\sum_{j=0}^{i-1}(d_i-d_j)^2 x_i y_j\right)\\ &\geq\left(\sum_{l=0}^{n} y_l\right)^2\left(\sum_{i=0}^{n}\sum_{j=0}^{i-1}(d_i-d_j)^2 x_i x_j\right)+\left(\sum_{l=0}^{n} x_l\right)^2\left(\sum_{i=0}^{n}\sum_{j=0}^{i-1}(d_i-d_j)^2 y_i y_j\right).\end{aligned} \tag{6.35}$$

Create the symmetric matrix D which has elements $D_{ij} = (d_i - d_j)^2$, and the vectors $X = (x_0, \ldots, x_n)$ and $Y = (y_0, \ldots, y_n)$. Multiply the expression (6.35) on each side by two and divide each side by $(\sum_{l=0}^{n} x_l)^2(\sum_{l=0}^{n} y_l)^2$. Expression (6.35) is then equivalent to the following:

$$\left(\frac{X}{\sum_{l=0}^{n} x_l}-\frac{Y}{\sum_{l=0}^{n} y_l}\right) D \left(\frac{X}{\sum_{l=0}^{n} x_l}-\frac{Y}{\sum_{l=0}^{n} y_l}\right)' \leq 0. \tag{6.36}$$

Equation (6.36) has the form

$$WDW' \leq 0$$

where $\sum_{l=0}^{n} w_l = 0$. The result follows from Lemma 6.7. □

Theorem 6.3. *The integrands R_2 and R_4 are concave functions of π for all values of y, and, as a consequence, the terms R_2 and R_4 are concave functions of π.*

Proof. We apply Lemma 3.3 to the integrand of R_2. The proof for the integrand of R_4 follows exactly the same steps.

Consider the integrand of R_2 in the form of expression (6.28), where again we take $g(k) = E(\mu_1|y, \tau_x = k)$. That is,

$$\sum_{k=0}^{n}\sum_{l=0}^{k-1}(g(k)-g(l))^2\frac{f(y|\tau_x = l)\pi_l f(y|\tau_x = k)\pi_k}{\sum_{r=0}^{n} f(y|\tau_x = r)\pi_r}.$$

Let $\pi_{(a)}$ and $\pi_{(b)}$ be the barycentric coordinates of two points in S^n, and let $\pi_{(a),k}$ and

$\pi_{(b),k}$ the kth components of $\pi_{(a)}$ and $\pi_{(b)}$ respectively. We want to show that

$$\begin{aligned}
&\sum_{k=0}^{n}\sum_{l=0}^{k-1}(g(k)-g(l))^2\frac{f(y|\tau_x=l)(t_a\pi_{(a),l}+t_b\pi_{(b),l})f(y|\tau_x=k)(t_a\pi_{(a),k}+t_b\pi_{(b),k})}{\sum_{r=0}^{n}t_af(y|\tau_x=r)\pi_{(a),r}+\sum_{s=0}^{n}t_bf(y|\tau_x=s)\pi_{(b),s}}\\
&\geq t_a\sum_{k=0}^{n}\sum_{l=0}^{k-1}(g(k)-g(l))^2\frac{f(y|\tau_x=l)\pi_{(a),l}f(y|\tau_x=k)\pi_{(a),k}}{\sum_{r=0}^{n}f(y|\tau_x=r)\pi_{(a),r}}\\
&+t_b\sum_{k=0}^{n}\sum_{l=0}^{k-1}(g(k)-g(l))^2\frac{f(y|\tau_x=l)\pi_{(b),l}f(y|\tau_x=k)\pi_{(b),k}}{\sum_{s=0}^{n}f(y|\tau_x=s)\pi_{(b),s}}
\end{aligned} \tag{6.37}$$

Rearranging, we find the above inequality to be equivalent to

$$\begin{aligned}
&\sum_{k=0}^{n}\sum_{l=0}^{k-1}(g(k)-g(l))^2\\
&\times\frac{(t_af(y|\tau_x=l)\pi_{(a),l}+t_bf(y|\tau_x=l)\pi_{(b),l})(t_af(y|\tau_x=k)\pi_{(a),k}+t_bf(y|\tau_x=k)\pi_{(b),k})}{\sum_{r=0}^{n}t_af(y|\tau_x=r)\pi_{(a),r}+\sum_{s=0}^{n}t_bf(y|\tau_x=s)\pi_{(b),s}}\\
&\geq\sum_{k=0}^{n}\sum_{l=0}^{k-1}(g(k)-g(l))^2\frac{t_a^2f(y|\tau_x=l)\pi_{(a),l}f(y|\tau_x=k)\pi_{(a),k}}{\sum_{r=0}^{n}t_af(y|\tau_x=r)\pi_{(a),r}}\\
&+\sum_{k=0}^{n}\sum_{l=0}^{k-1}(g(k)-g(l))^2\frac{t_b^2f(y|\tau_x=l)\pi_{(b),l}f(y|\tau_x=k)\pi_{(b),k}}{\sum_{s=0}^{n}t_2f(y|\tau_x=s)\pi_{(b),s}}.
\end{aligned}$$

Setting,

$$\begin{aligned}
d_k &= g(k)\\
d_l &= g(l)\\
X_k &= t_af(y|\tau_x=k)\pi_{(a),k}\\
X_l &= t_af(y|\tau_x=l)\pi_{(a),l}\\
Y_k &= t_bf(y|\tau_x=k)\pi_{(b),k}\\
Y_l &= t_bf(y|\tau_x=l)\pi_{(b),l}
\end{aligned}$$

we have

$$\begin{aligned}
&\sum_{k=0}^{n}\sum_{l=0}^{k-1}(d_k - d_l)^2 \frac{(X_l + Y_l)(X_k + Y_k)}{(\sum_{r=0}^{n} X_r + \sum_{s=0}^{n} Y_s)} \\
&\geq \sum_{k=0}^{n}\sum_{l=0}^{k-1}(d_k - d_l)^2 \frac{X_k X_l}{\sum_{r=0}^{n} X_r} \\
&+ \sum_{k=0}^{n}\sum_{l=0}^{k-1}(d_k - d_l)^2 \frac{Y_k Y_l}{\sum_{s=0}^{n} Y_s}.
\end{aligned} \tag{6.38}$$

Note that $g(k)$, and, hence d_k, is in $\mathbb{R}$, and that the quantities X_k and Y_k are all positive. Thus, by Lemma 6.8, the inequality (6.38) is satisfied, and hence the integrand of R_2 is concave. The proof for R_4 follows in exactly the same way. □

Combining our results for the terms R_1, R_3, R_2, and R_4, we have Theorem 6.4.

Theorem 6.4. *Suppose we have a single-path problem with a log-concave prior distribution for the changepoint. Then the optimal design for estimating the before-and-after-change means μ_1 and μ_2, when using a Bayes risk based on squared error loss, is one of the designs in V.*

Proof. From Lemma 6.5 we have that R_1 and R_3 are linear in π. From Theorem 6.3 we have that R_2 and R_4 are concave in π. Hence their sum R, the Bayes risk based on squared error loss, is concave in π. From Theorem 5.6 we know that a concave design criterion is minimized at one of the vertices in V. □

Next we observe that, if we allowed design points to crowd together, we would minimize over the set of design measures occupying all of S^n. By inspection, we see that R_2 and R_4 are zero at the vertices of S^n. This is because by placing all the points at 0 and T we keep the posterior expectations $E(\mu_1|y, \tau_x = k)$ and $E(\mu_2|y, \tau_x = k)$ constant in the interval $[0, T]$ and hence they have zero variance in the observation interval $[0, T]$. The implication is that if the designs points were allowed to crowd together, then the optimal design for the Bayes risk based on a squared error loss would be given by R_1+R_3. The optimal design from the generalized Lauter's criterion and the Bayes risk based on squared error loss would then be the same. We will see in Chapter 7, that when the difference between the hyperparameter means is large compared to the hyperparametric and model variances, the generalized Lauter's criterion is a good approximation to the Bayes risk based on squared error loss.

Before concluding this section, we note that, in changepoint analysis, researchers sometimes prefer to make inference about the difference in means, rather than about μ_1 and μ_2 separately. In this setting the obvious parameters would be μ and $\mu+\delta$. The inference is then about δ. Using techniques similar to the ones in this section, we can show a Bayes risk based on a squared error loss for δ,

$$\int Var(\delta|y)f(y)dy, \tag{6.39}$$

is also concave in π. Theorem 5.6 can then be used to show that the optimal design for estimating δ will be one of the designs in V.

6.3.2 Design Criterion Function for Estimation the Changepoint Location

One can also use the Bayes risk based on squared error loss as a design criterion function for estimating the changepoint location τ. Unfortunately, we can not use our design measure π to provide general optimal design results for estimating the changepoint location. Our design measure arose by combining the random variable τ with the design x to form the discrete random variable τ_x. Any design criterion function for estimating the changepoint will have to be expressed in terms of τ directly. Generally these design criterion functions will lead to a non-linear optimization problem.

For the sake of illustration we provide the Bayes risk based on squared error loss for estimating the changepoint below.

$$\int Var(\tau|y)f(y)dy \tag{6.40}$$

6.4 Combining Criterion Functions

As we mentioned in Section 1.2.2, design criterion functions, with testing and estimation respectively as the ultimate goals, can be combined to obtain a single design criterion function. Here, in fact, we have three criterion functions (two for testing and one for estimating) which are all concave in π. By taking a convex combination of one of the testing criterion functions with the estimation criterion function, we obtain a criterion function that is again concave in π. In the convex combination, one can weight the two criterion functions according to the relative importance attached to the two problems. For instance, if testing is more important, we can put more weight on the testing criterion function.

Theorem 6.5. *Consider the single-path problem with a log-concave prior distribution for the changepoint. Then*
A) Suppose that a criterion function is constructed by taking a convex combination of the generalized 0-1 Bayes risk (for testing for a change) and the squared error loss Bayes risk (for estimation of the means). Then fixing x_n at T, the optimal design is one of the vectors $\{u_1, \ldots, u_n\}$.
B) Suppose that a criterion function is constructed by taking a convex combination of the Spezzaferri criterion function (for testing for a change) and the squared error loss Bayes risk (for estimation of the means). Then fixing x_n at T, the optimal design is one of the vectors $\{u_1, \ldots, u_n\}$.

Proof. The proof of A) and B) follow directly from Theorem 5.7 since any convex combination of design criterion functions which are concave in π is concave in π. □

Theorem 6.6. *Given C and D in (5.14) and (5.15), consider the single-path problem with a log-concave prior distribution for the changepoint. Then*
A) Suppose that a criterion function is constructed by taking a convex combination of the generalized 0-1 Bayes risk (for testing for a change in $[t_1, t_2]$) and the squared error loss Bayes risk (for estimation of the means). Then fixing points x_q and x_{q+1} at t_1 and t_2 respectivley, the optimal design is a Cartesian product of an element of C with an element of D.
B) Suppose that a criterion function is constructed by taking a convex combination of the Spezzaferri criterion function (for testing for a change in $[t_1, t_2]$) and the squared error loss Bayes risk (for estimation of the means). Then fixing points x_q and x_{q+1} at t_1 and t_2 respectivley, the optimal design is a Cartesian product of an element of C with an element of D.

Proof. The proof of A) and B) follow directly from Theorem 5.8 since any convex combination of design criterion functions which are concave in π is concave in π. □

In the next chapter we consider two particular single-path models.

Chapter 7

Particular Single-Path Changepoint Models

We now know that the optimal designs for the problems discussed in Chapter 6 lie amongst a manageable number of designs. Nevertheless we must still compute the design criterion function for each design in the set, to determine which of the designs is optimal. To make concrete the rather abstract discussion of the previous chapter, we consider two particular single-path changepoint models and simulations of the Bayes risk based on squared error loss for estimating the before-and-after-change means.

The first model considered in Section 7.1 arises from measurements taken a distance d apart and assumed to be conditionally independent. The data are very general as we allow them to have any NEF distribution. We parametrize the model in terms of the canonical parameters θ_1 and θ_2. The designs we consider are optimal for estimating the means $K'_\eta(\theta_1)$ and $K'_\eta(\theta_2)$. We use DY-conjugate prior distributions for the before-and-after-change canonical parameters θ_1 and θ_2. By using DY-conjugate prior distributions we are able to compute the terms R_1 and R_3 analytically.

The second model we consider is the common changepoint multi-path problem. In this problem we have an arbitrary number of subjects m, and we assume that all subjects change at exactly the same time (that is, there is exactly one changepoint). In Chapter 8, we consider the multiple changepoints multi-path problem, where each subject has his or her own changepoint. Although not as realistic as the multiple changepoints problem, the common changepoint problem seems to be a good approximation to the multiple changepoints problem when the variance of the subject changepoints is small. In the multi-path models, each subject has random effect before-change and after-change means normally distributed about distinct hierarchical means.

Our simulations for the common changepoint multi-path problem consist of the Bayes risk based on squared error loss for estimation of the before-and-after-change hierarchical means. Due to the constraint that all subjects change at the same time, the common changepoint multi-path problem can be treated as a single-path changepoint problem by taking the average of the measurements at each design point. Depending on the changepoint location and the design, each average of m measurements is then normally distributed about either the before-change or the after-change hierarchical mean. The random effects induce a correlation between the averages; however since this correlation does not depend on the design, our earlier optimal design results for the single-path problem follow. By using conjugate normal prior distributions for the before-and-after-change hierachical means, we can calculate the terms R_1 and R_3 analytically.

In the simulations of each model, we observe that the generalized Lauter's criterion function $R_1 + R_3$ is a good approximation of the Bayes risk based on squared error loss when the difference in the hyperparametric means is large compared to the model variances and the hyperparametric variances.

7.1 The NEF Single-Path Changepoint Problem

Let the NEF member $p(y|\theta_1) = \exp(\theta_1 y - K_\eta(\theta_1))\eta(y)$ be the before-change distribution of our data and the NEF member $p(y|\theta_2) = \exp(\theta_2 y - K_\eta(\theta_2))\eta(y)$ be the after-change distribution. Hence the before-change mean is $K'_\eta(\theta_1)$ and the after-change mean is $K'_\eta(\theta_2)$.

We use the DY-conjugate prior densities in (7.1) and (7.2) for the canonical parameters θ_1 and θ_2:

$$p_{\nu_1,\lambda_1}(\theta_1) = M(\nu_1, \lambda_1) \exp(\nu_1\theta_1 - \lambda_1 K_\eta(\theta_1)) \tag{7.1}$$

$$p_{\nu_2,\lambda_2}(\theta_2) = M(\nu_2, \lambda_2) \exp(\nu_2\theta_2 - \lambda_2 K_\eta(\theta_2)) \tag{7.2}$$

The unknown parameters θ_1, θ_2 and τ are assumed to be independent.

As in Section 4.1, let y_i be the measurement taken at the x_ith design point. Again, the vector $y = (y_1, \ldots, y_n)$ comprises the measurements taken at the design points $(x_1, \ldots, x_n)$. Rewriting the likelihood (4.1) in (7.3), we have that the measurements y are conditionally independent given the canonical parameters θ_1 and θ_2, and the changepoint τ. We then have

$$p(y|\theta_1, \theta_2, \tau) = \prod_{x_i \leq \tau} p(y_i|\theta_1) \prod_{x_i > \tau} p(y_i|\theta_2). \tag{7.3}$$

Substituting the NEF densities into the likelihood (7.3), we see that the likelihood (7.3) is immediately expressible in terms of the random variable τ_x:

$$\begin{aligned} p(y|\theta_1, \theta_2, \tau_x = k) = \exp & \left(\theta_1 \sum_{i \leq k} y_i - kK_\eta(\theta_1)\right) \\ & \times \exp\left(\theta_2 \sum_{i>k} y_i - (n-k)K_\eta(\theta_2)\right) \prod_{i \leq k} \eta(y_i) \prod_{i>k} \eta(y_i). \end{aligned} \tag{7.4}$$

Next, we present the posterior densities calculated from the likelihood (7.4), the DY-conjugate prior densities (7.1) and (7.2), and the prior density $f(\tau)$. These posterior densities are necessary for calculating our estimators $E(K'_\eta(\theta_1)|y)$ and $E(K'_\eta(\theta_2)|y)$ and for evaluating the Bayes risk based on squared error loss.

The calculation of y given $\tau_x = k$ is quite simple and proceeds as follows:

$$\begin{aligned} p(y|\tau_x = k) &= \int\int p(y, \theta_1, \theta_2|\tau_x = k) d\theta_1 d\theta_2 \\ &= \int\int p(y|\theta_1, \theta_2, \tau_x = k) p_{\nu_1,\lambda_1}(\theta_1) p_{\nu_2,\lambda_2}(\theta_2) d\theta_1 d\theta_2 \\ &= \frac{M(\nu_1, \lambda_1) M(\nu_2, \lambda_2) \prod_{i\leq k} \eta(y_i) \prod_{i>k} \eta(y_i)}{M(\nu_1 + \sum_{i\leq k} y_i, \lambda_1 + k) M(\nu_2 + \sum_{i>k} y_i, \lambda_2 + (n-k))}. \end{aligned} \tag{7.5}$$

The density $p(y|\tau_x = k)$ can then be used to obtain the posterior densities $p(\theta_1|y, \tau_x = k)$ and $p(\theta_2|y, \tau_x = k)$:

$$\begin{aligned} p(\theta_1|y, \tau_x = k) &= \frac{\int p(y, \theta_1, \theta_2, \tau_x = k) d\theta_2}{p(y|\tau_x = k)\pi_k} \\ &= \frac{\int p(y, \theta_1, \theta_2|\tau_x = k) d\theta_2}{p(y|\tau_x = k)} \\ &= M\left(\nu_1 + \sum_{i\leq k} y_i, \lambda_1 + k\right) \\ &\times \exp\left(\left(\nu_1 + \sum_{i\leq k} y_i\right)\theta_1 - (\lambda_1 + k) K_\eta(\theta_1)\right). \end{aligned} \tag{7.6}$$

The calculation for $p(\theta_2|y, \tau_x = k)$ is identical, with θ_1 and θ_2 interchanged and yields

$$\begin{aligned} p(\theta_2|y, \tau_x = k) = M&\left(\nu_2 + \sum_{i>k} y_i, \lambda_2 + (n-k)\right) \\ \times \exp&\left(\left(\nu_2 + \sum_{i>k} y_i\right)\theta_2 - (\lambda_2 + (n-k)) K_\eta(\theta_2)\right). \end{aligned} \tag{7.7}$$

Naturally, the posterior densities for θ_1 and θ_2, (7.6) and (7.7) have the same form as the standard conjugate prior distributions for θ_1 and θ_2, but with the hyperparameters ν_1, ν_2, λ_1, and λ_2 updated by the data.

The marginal distribution of y is calculated below and is a mixture distribution of the posterior densities $p(y|\tau_x = k)$ with the weights π_k:

$$\begin{aligned} p(y) &= \sum_{k=0}^{n} p(y|\tau_x = k)\pi_k \\ &= \sum_{k=0}^{n} \frac{M(\nu_1, \lambda_1)M(\nu_2, \lambda_2)\prod_{i\leq k}\eta(y_i)\prod_{i>k}\eta(y_i)}{M\left(\nu_1 + \sum_{i\leq k} y_i, \lambda_1 + k\right) M\left(\nu_2 + \sum_{i>k} y_i, \lambda_2 + (n-k)\right)}\pi_k. \end{aligned} \tag{7.8}$$

Using the marginal distribution of y given in (7.8) we find

$$\begin{aligned} p(\tau_x = k|y) &= \frac{p(y|\tau_x = k)\pi_k}{\sum_{l=0}^{n} p(y|\tau_x = l)\pi_l} \\ &= \frac{\frac{M(\nu_1,\lambda_1)M(\nu_2,\lambda_2)\prod_{i\leq k}\eta(y_i)\prod_{i>k}\eta(y_i)}{M(\nu_1+\sum_{i\leq k} y_i,\lambda_1+k)M(\nu_2+\sum_{i>k} y_i,\lambda_2+(n-k))}\pi_k}{\sum_{l=0}^{n} \frac{M(\nu_1,\lambda_1)M(\nu_2,\lambda_2)\prod_{i\leq l}\eta(y_i)\prod_{i>l}\eta(y_i)}{M(\nu_1+\sum_{i\leq l} y_i,\lambda_1+l)M(\nu_2+\sum_{i>l} y_i,\lambda_2+(n-l))}\pi_l}. \end{aligned} \tag{7.9}$$

We finish by presenting, in Theorem 7.1, the posterior expectations of the means, $K'_\eta(\theta_1)$ and $K'_\eta(\theta_2)$, given y and $\tau_x = k$. These posterior expectations play a key role when evaluating the Bayes risk. Note that the DY-conjugate prior distributions lead to posterior expectations $E(K'_\eta(\theta_1)|y, \tau_x = k)$ and $E(K'_\eta(\theta_2)|y, \tau_x = k)$ that are linear functions of y.

Theorem 7.1. *For the likelihood (7.4) and DY-conjugate prior densities (7.6) and (7.7)*

$$E(K'_\eta(\theta_1)|y, \tau_x = k) = \frac{\nu_1 + \sum_{i\leq k} y_i}{\lambda_1 + k} \quad \textit{and} \quad E(K'_\eta(\theta_2)|y, \tau_x = k) = \frac{\nu_2 + \sum_{i>k} y_i}{\lambda_2 + (n-k)}.$$

Proof. We present the proof for $E(K'_\eta(\theta_1)|y, \tau_x = k)$. The proof for $E(K'_\eta(\theta_2)|y, \tau_x = k)$ is exactly the same.

Since $p(\theta_1|y, \tau_x = k)$, in (7.6), is a density, we have $\int p(\theta_1|y, \tau_x = k)d\theta_1 = 1$. Differentiating with respect to θ_1 we obtain $\frac{d}{d\theta_1}\int p(\theta_1|y, \tau_x = k)d\theta_1 = 0$. By carrying the derivative under the integral sign we obtain the desired result. □

7.1.1 The Bayes Risk Based on Squared Error Loss

We perform the same operations as in Section 6.3.1 to re-express the Bayes risk in terms of R_1, R_2, R_3, and R_4. The only difference here is that the before-and-after-change means and the densities used in the calculations are expressed in terms of the canonical parameters θ_1 and θ_2.

Below is the loss as a function of the canonical parameters; our loss has two terms since we are estimating two means:

$$(K'_\eta(\theta_1) - E(K'_\eta(\theta_1)|y))^2 + (K'_\eta(\theta_2) - E(K'_\eta(\theta_2)|y))^2. \tag{7.10}$$

To obtain the Bayes risk R, we integrate with respect to $p(y|\theta_1, \theta_2)$ and $p(\theta_1)p(\theta_2)$.

$$\begin{aligned} R = & \int\int\int (K'_\eta(\theta_1) - E(K'_\eta(\theta_1)|y))^2 p(y|\theta_1,\theta_2)p(\theta_1)p(\theta_2)dyd\theta_1 d\theta_2 \\ + & \int\int\int (K'_\eta(\theta_2) - E(K'_\eta(\theta_2)|y))^2 p(y|\theta_1,\theta_2)p(\theta_1)p(\theta_2)dyd\theta_1 d\theta_2 \end{aligned}$$

Again, we simplify to get

$$R = \int Var(K'_\eta(\theta_1)|y)p(y)dy + \int Var(K'_\eta(\theta_2)|y)p(y)dy. \tag{7.11}$$

In the same way identity (6.22) was used in Chapter 6, we now use identity (7.12) to divide expression (7.11) into four terms. In terms of the canonical parameters, we have

$$Var(K'_\eta(\theta)|y) = E_{\tau_x = k|y}(Var(K'_\eta(\theta)|y, \tau_x = k)) + Var_{\tau_x = k|y}(E(K'_\eta(\theta)|y, \tau_x = k)). \tag{7.12}$$

The term R re-expressed as the sum of R_1, R_2, R_3, and R_4 is

$$\begin{aligned} R = & \int E_{\tau_x = k|y}(Var(K'_\eta(\theta_1)|y, \tau_x = k))p(y)dy \\ + & \int Var_{\tau_x = k|y}(E(K'_\eta(\theta_1)|y, \tau_x = k))p(y)dy \\ + & \int E_{\tau_x = k|y}(Var(K'_\eta(\theta_2)|y, \tau_x = k))p(y)dy \\ + & \int Var_{\tau_x = k|y}(E(K'_\eta(\theta_2)|y, \tau_x = k))p(y)dy. \end{aligned} \tag{7.13}$$

Again the terms R_1 and R_3 can be integrated, while the terms R_2 and R_4 lead to a non-linear integral in y.

By Fubini's theorem R_1 and R_3 can be re-expressed as:

$$R_1 = \sum_{k=0}^{n} E_{y|\tau_x=k}(Var(K'_\eta(\theta_1)|y, \tau_x = k))\pi_k \tag{7.14}$$

and

$$R_3 = \sum_{k=0}^{n} E_{y|\tau_x=k}(Var(K'_\eta(\theta_2)|y, \tau_x = k))\pi_k. \tag{7.15}$$

Since the calculations for R_1 and R_3 are very similar, we evaluate the expression for $E_{y|\tau_x=k}(Var(K'_\eta(\theta_1)|y, \tau_x = k))$ in detail, and comment on the result for $E_{y|\tau_x=k}(Var(K'_\eta(\theta_2)|y, \tau_x = k))$. Before finding $E_{y|\tau_x=k}(Var(K'_\eta(\theta_1)|y, \tau_x = k))$, we give two Lemmas.

Lemma 7.1. *Given the model described by the likelihood (7.4), and the prior densities (7.1) and (7.2), we have*

$$Var(K'_\eta(\theta_1)|\tau_x = k) = \frac{1}{\lambda_1}E(K''_\eta(\theta_1)) \quad \textit{and} \quad Var(K'_\eta(\theta_2)|\tau_x = k) = \frac{1}{\lambda_2}E(K''_\eta(\theta_2)).$$

Proof. We present the proof of $Var(K'_\eta(\theta_1)|\tau_x = k) = \frac{1}{\lambda_1}E(K''_\eta(\theta_1))$. The proof for the equality $Var(K'_\eta(\theta_2)|\tau_x = k) = \frac{1}{\lambda_2}E(K''_\eta(\theta_2))$ is exactly the same.

We have assumed in our model that θ_1 and τ are independent. Since τ_x is a function of τ and the design x, which is not a random variable, we have that θ_1 is also independent of τ_x. Therefore $Var(K'_\eta(\theta_1)|\tau_x = k) = Var(K'_\eta(\theta_1))$. Now to show $Var(K'_\eta(\theta_1)|\tau_x = k) = \frac{1}{\lambda_1}E(K''_\eta(\theta_1))$ we differentiate the equation $\int p_{\nu_1,\lambda_1}(\theta_1)d\theta_1 = 1$ twice with respect to θ_1. That is, we carry the second derivative under the integral and solve the equation $\frac{d^2}{d\theta_1^2}\int p_{\nu_1,\lambda_1}(\theta_1)d\theta_1 = 0$. □

Lemma 7.2. *Given the model described by the likelihood (7.4) and the prior densities (7.1) and (7.2), we have*

$$Var_{y|\tau_x=k}\left(\sum_{i\leq k} y_i\right) = k(k+\lambda_1)Var(K'_\eta(\theta_1))$$

and

$$Var_{y|\tau_x=k}\left(\sum_{i>k} y_i\right) = (n-k)((n-k)+\lambda_2)Var(K'_\eta(\theta_2)).$$

Proof. We prove the result for $Var_{y|\tau_x=k}(\sum_{i\leq k} y_i)$. The calculation for $Var_{y|\tau_x=k}(\sum_{i>k} y_i)$ is exactly the same. We have the well-known identity

$$Var_{y|\tau_x=k}\left(\sum_{i\leq k} y_i\right) = E_{\theta_1}\left(Var_{y|\tau_x=k,\theta_1}\left(\sum_{i\leq k} y_i\right)\right) + Var_{\theta_1}\left(E_{y|\tau_x=k,\theta_1}\left(\sum_{i\leq k} y_i\right)\right).$$

Recalling that since the y_i's, for $i \leq k$, are independent and identically distributed given θ_1, θ_2 and $\tau_x = k$, the above identity simplifies to

$$Var_{y|\tau_x=k}\left(\sum_{i\leq k} y_i\right) = E_{\theta_1}(kVar_{y|\tau_x=k,\theta_1}(y_j)) + Var_{\theta_1}(kE_{y|\tau_x=k,\theta_1}(y_j)),$$

where j is any j less than or equal to k.

Hence, we have,

$$Var_{y|\tau_x=k}\left(\sum_{i\leq k} y_i\right) = kE_{\theta_1}(K''_\eta(\theta_1)) + k^2Var_{\theta_1}(K'_\eta(\theta_1)).$$

Using Lemma 7.1, we have

$$Var_{y|\tau_x=k}\left(\sum_{i\leq k} y_i\right) = k\lambda_1 Var_{\theta_1}(K'_\eta(\theta_1)) + k^2Var_{\theta_1}(K'_\eta(\theta_1)). \tag{7.16}$$

Upon simplification of (7.16), we obtain the desired result. To show

$$Var_{y|\tau_x=k}\left(\sum_{i>k} y_i\right) = (n-k)\lambda_2 Var_{\theta_2}(K'_\eta(\theta_2)) + (n-k)^2Var_{\theta_2}(K'_\eta(\theta_2)), \tag{7.17}$$

we follow exactly the same steps. Of course, θ_2, μ_2 and λ_2 are interchanged with θ_1, μ_1 and λ_1 and $(n-k)$ appears in the equation instead of k, because there are $(n-k)$ identically distributed $y_i's$ with $i > k$. □

Theorem 7.2. *Given the model described by the likelihood (7.4) and the prior densities (7.1) and (7.2), we have* $E_{y|\tau_x=k}(Var(K'(\theta_1)|y,\tau_x=k)) = \frac{\lambda_1}{\lambda_1+k}Var(K'_\eta(\theta_1))$ *and* $E_{y|\tau_x=k}(Var(K'(\theta_2)|y,\tau_x=k)) = \frac{\lambda_2}{\lambda_2+(n-k)}Var(K'_\eta(\theta_2))$.

Proof. Again, since the derivation of $E_{y|\tau_x=k}(Var(K'(\theta_2)|y,\tau_x = k))$ mimics that of $E_{y|\tau_x=k}(Var(K'(\theta_1)|y,\tau_x = k))$, we present the derivation of $E_{y|\tau_x=k}(Var(K'(\theta_1)|y,\tau_x = k))$.

Using the identity,

$$\begin{aligned}&Var(K'_\eta(\theta_1)|\tau_x = k)\\ &= E_{y|\tau_x=k}(Var(K'_\eta(\theta_1)|y,\tau_x = k)) + Var_{y|\tau_x=k}(E(K'_\eta(\theta_1)|y,\tau_x = k)),\end{aligned}$$

we have

$$\begin{aligned}&E_{y|\tau_x=k}(Var(K'_\eta(\theta_1)|y,\tau_x = k))\\ &= Var(K'_\eta(\theta_1)|\tau_x = k) - Var_{y|\tau_x=k}(E(K'_\eta(\theta_1)|y,\tau_x = k)).\end{aligned} \tag{7.18}$$

Recalling that θ_1 and τ_x are independent, and substituting $E(K'_\eta(\theta_1)|y,\tau_x = k)$ from Theorem 7.1, equation (7.18) becomes

$$\begin{aligned}&E_{y|\tau_x=k}(Var(k'_\eta(\theta_1)|y,\tau_x = k))\\ &= Var(k'_\eta(\theta_1)) - \left(\frac{1}{\lambda_1 + k}\right)^2 Var_{y|\tau_x=k}\left(\sum_{i\leq k} y_i\right).\end{aligned} \tag{7.19}$$

Using Lemma 7.2 and simplifying, we obtain the desired result. Following the same steps, we get a similar expression for $E_{y|\tau_x=k}(Var(K'_\eta(\theta_2)|y,\tau_x = k))$ with θ_2, λ_2, and ν_2 interchanged with θ_1, λ_1, and ν_1 and $(n-k)$ interchanged with k. □

We now have an analytical expression for $R_1 + R_3$. This expression is of the form $\sum_{k=0}^n H(k)\pi_k$ where $H(k) = \frac{\lambda_1}{\lambda_1+k}Var(K'_\eta(\theta_1)) + \frac{\lambda_2}{\lambda_2+(n-k)}Var(K'_\eta(\theta_2))$.

Example 7.1. $H(k)$ **for Normal Data**

If our data points y_i have a normal distribution, then, as seen in Example 3.9, the DY-conjugate prior distributions for the before-and-after-change canonical parameters are equivalent to normal prior distributions for the before-and-after-change means, which we will denote by μ_1 and μ_2. Denoting the hyperparameters for μ_1 as $\bar{\mu}_1$ and $\bar{\sigma}_1$ and *the hyperparameters for* μ_2 *as* $\bar{\mu}_2$ *and* $\bar{\sigma}_2$ *we have* $\lambda_1 = \frac{\sigma_1^2}{\bar{\sigma}_1^2}$, *and* $\lambda_2 = \frac{\sigma_2^2}{\bar{\sigma}_2^2}$. *Furthermore, it is obvious that* $Var(K_\eta'(\theta_1)) = \bar{\sigma}_1^2$ *and* $Var(K_\eta'(\theta_2)) = \bar{\sigma}_2^2$. *Therefore,*

$$H(k) = \frac{1}{\frac{1}{\bar{\sigma}_1^2} + \frac{k}{\sigma_1^2}} + \frac{1}{\frac{1}{\bar{\sigma}_2^2} + \frac{n-k}{\sigma_2^2}}. \tag{7.20}$$

Analytical expressions for R_2 and R_4 are, of course, unavailable. However, from Section 6.3.1 we know these two terms are concave over S^n. In the next section we present numerical examples of the Bayes risk based on squared error loss for normally distributed data.

7.1.2 Simulations

We present three numerical examples based on Normal data. The first two examples, have two design points. We use only two design points because the dimension of the problem is small enough to easily plot the risk. Examples 7.2 and 7.3, illustrate the situation where the Bayes risk can be approximated by $R_1 + R_3$, the generalized Lauter criterion function, under estimation of the before-and-after-means. This situation arises when the differences in hyperparametric means is large relative to the hyperparametric and model variances. Under these circumstances, it can be shown that the integrands of R_2 and R_4 remain small. This observation is not enough to prove that R_2 and R_4 will be small when the means are far apart compared to the variances but it is suggestive that this is the case. In Example 7.4, we consider a more realistic situation with five design points. All three of our examples demonstrate that the optimal design is ultimately a function of the hyperparameters and the changepoint prior distribution.

We begin in Example 7.2 with model variances and hyperparameters such that the difference in hyperparametric means is small compared to the hyperparametric variances and model variances. In Example 7.3 we consider the opposite situation, where the model and hyperparametric variances are small compared to the difference in the hyperparametric means. For Example 7.4 we consider model and hyperparametric variances such that the before-change variances are equal to the after-change variances.

Example 7.2. Small Difference in Hyperparametric Means Compared to Variances

In Table 7.1 and Figures 7.1, 7.2 and 7.3, we present the results of numerical simulations of the Bayes risk for the single-path model with variances $\sigma_1^2 = 2.5$ *and* $\sigma_2^2 = 3$ *and hyperparameters* $\bar{\mu}_1 = 4.5$, $\bar{\mu}_2 = 4$, $\bar{\sigma}_1^2 = 3$ *and* $\bar{\sigma}_2^2 = 2$. *We need conjugate normal prior distributions for the before-and-after-change means and the truncated log-concave normal prior distribution of Figure 5.1 is used for the changepoint prior distribution. The length of the interval* T *is* 10 *and the minimum distance between design points, d, is 2.*

From Table 7.1 we see that the Bayes risk is minimized by the design u_1. *This table also shows us that* R_1 *is minimized by the design* u_0 *and that* R_3 *is minimized by the design* u_2. *In the sum* $R_1 + R_3$, R_1 *and* R_3 *compromise and the sum is minimized at the design* u_1.

Figures 7.1, 7.2 and 7.3 plot $R_1 + R_3$, $R_2 + R_4$ *and* R *respectively. We see the linear forms of* R_1 *and* R_3 *in Figure 7.1 and the concave form of* $R_2 + R_4$ *in 7.2. Finally, in Figure 7.3 we see the full concave Bayes risk* R *for these values of the model variances and hyperparameters.*

Example 7.3. Large Difference in Hyperparametric Means Compared to Variances

In Table 7.2 and Figures 7.4, 7.5 and 7.6 we present numerical simulations of the Bayes risk for the single-path model with variances $\sigma_1^2 = 1.5$ *and* $\sigma_2^2 = 1$, *and hyperparameters* $\bar{\mu}_1 = 19$, $\bar{\mu}_2 = 1$, $\bar{\sigma}_1^2 = 1$ *and* $\bar{\sigma}_2^2 = 2$. *We used conjugate normal prior distributions for the before-and-after-change means and the truncated log-concave nor-*

Design	R_1	R_3	$R_1 + R_3$	$R_2 + R_4$	R
u_0	0.955	1.923	2.883	0.140	3.023
u_1	1.397	1.193	2.590	0.038	2.628
u_2	2.766	0.906	3.672	0.219	3.891

Table 7.1: Values from the numerical simulation for the designs in the set V. Note that due to the numerical simulation of R_2+R_4, and the region $G_f(\mathbb{X}^n)$, these values are approximate. For example, clearly $R_2 + R_4$ should be 0 for u_1.

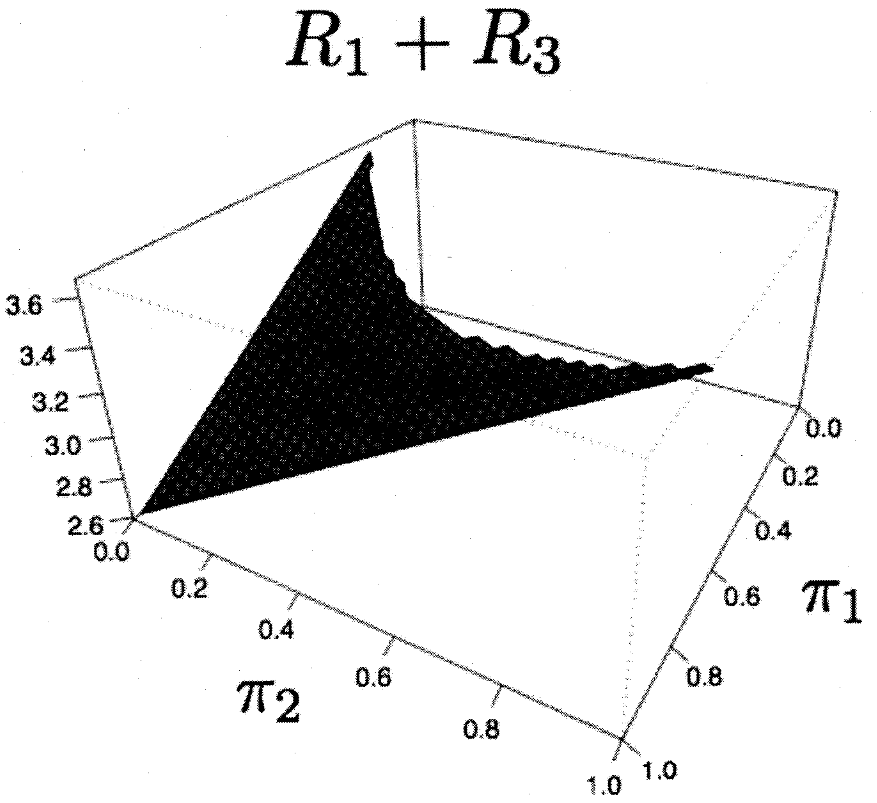


Figure 7.1: The terms R_1 and R_3 for the single-path model with hyperparameters $\bar{\mu}_1 = 4.5$, $\bar{\mu}_2 = 4$, $\bar{\sigma}_1^2 = 3$ and $\bar{\sigma}_2^2 = 2$ and variances $\sigma_1^2 = 2.5$ and $\sigma_2^2 = 3$.

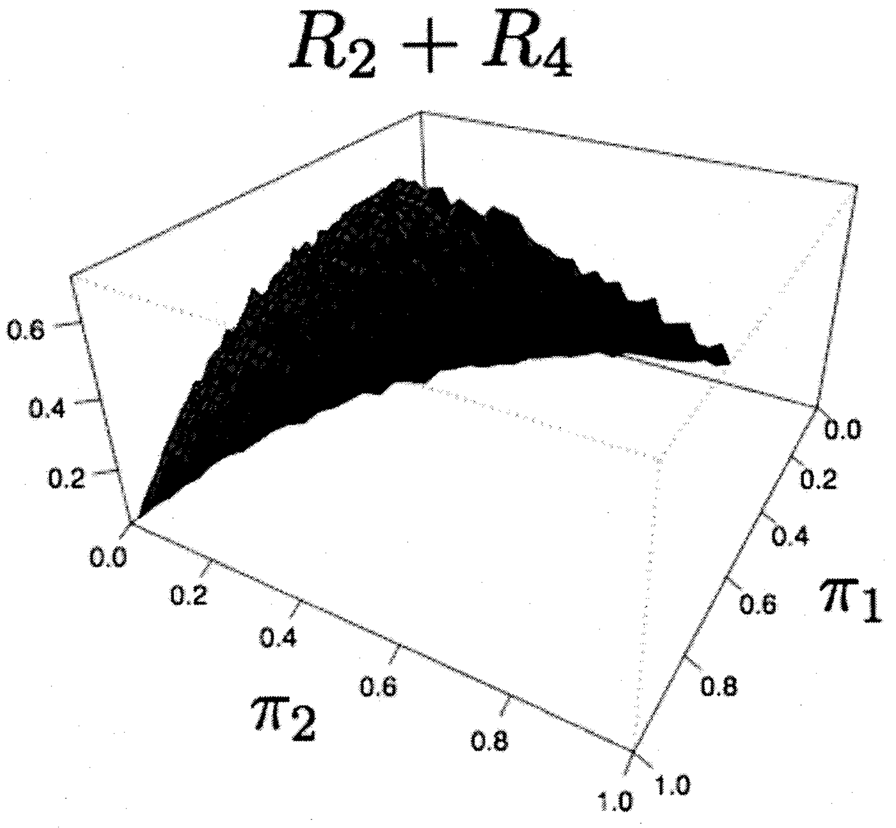


Figure 7.2: The terms R_2 and R_4 for the single-path model with hyperparameters $\bar{\mu}_1 = 4.5$, $\bar{\mu}_2 = 4$, $\bar{\sigma}_1^2 = 3$ and $\bar{\sigma}_2^2 = 2$ and variances $\sigma_1^2 = 2.5$ and $\sigma_2^2 = 3$.

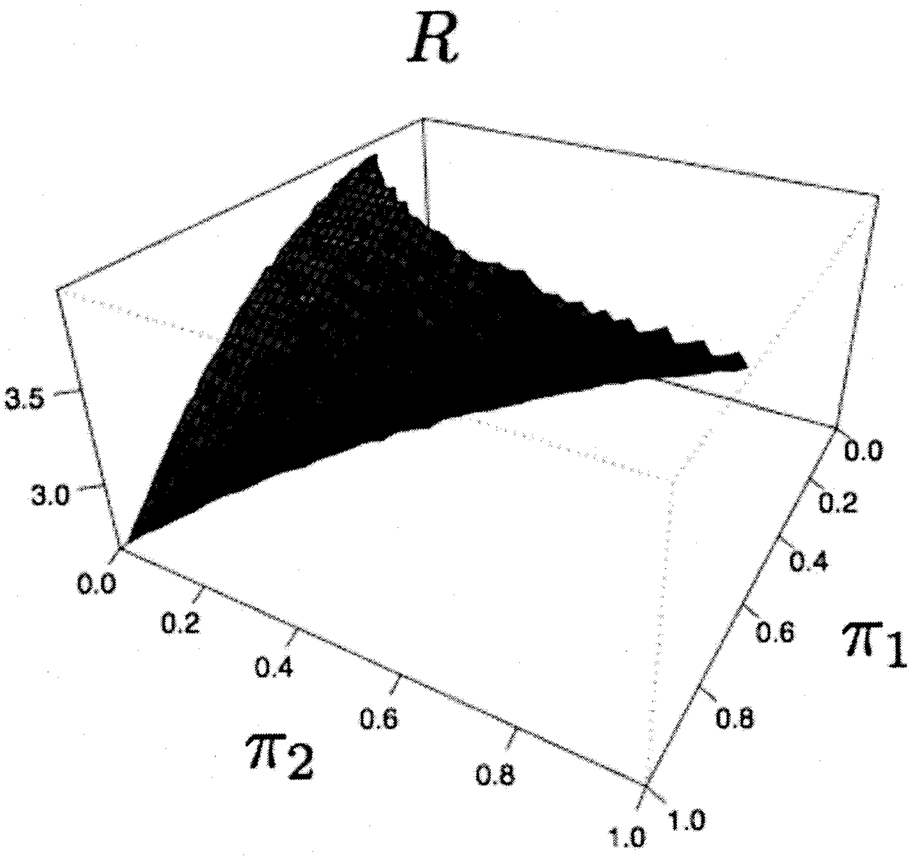


Figure 7.3: The Bayes risk based on squared error loss, R, for the single-path model with hyperparameters $\bar{\mu}_1 = 4.5$, $\bar{\mu}_2 = 4$, $\bar{\sigma}_1^2 = 3$ and $\bar{\sigma}_2^2 = 2$ and variances $\sigma_1^2 = 2.5$ and $\sigma_2^2 = 3$.

mal prior density of Figure 5.1 is used for the changepoint prior distribution. The length of the interval T is 10 *and the minimum distance between design points, d, is 2.*

From table 7.2 we see that the Bayes risk is minimized by the design u_1. We also see from this table that $R_2 + R_4$ is essentially zero over $G_f(\mathbb{X}^n)$. Again, R_1 is minimized by the design u_0 and R_3 is minimized by the design u_2. The sum $R_1 + R_3$ is minimized at the design u_1.

Figures 7.4, 7.5 and 7.6 plot $R_1 + R_3$, $R_2 + R_4$ and R, respectively. We see the linear structure of $R_1 + R_3$ in Figure 7.4 and that $R_2 + R_4$ are essentially zero in 7.5. Finally, in Figure 7.6 we see how $R_1 + R_3$ dominates the Bayes risk R, for these values of the model variances and hyperparameters.

Design	R_1	R_3	$R_1 + R_3$	$R_2 + R_4$	R
u_0	0.451	1.886	2.336	0.000	2.336
u_1	0.608	0.661	1.269	0.000	1.269
u_2	0.943	0.438	1.381	0.000	1.381

Table 7.2: Values from the numerical simulation for the designs in the set V.

Examples 7.2 and 7.3 demonstrate that when the difference in the hyperparametric means is large compared to the magnitudes of the variances, the Bayes risk is approximated by $R_1 + R_3$. That is, the generalized Lauter's criterion function and the Bayes risk based on the squared error loss are approximately equal in this situation.

Example 7.4. Five Design Points

To conclude, we present a numerical simulation with five design points. Here, our Bayes risk based on squared error loss criterion function for estimating the before-and-after-change means is a scalar function over a subset of a five-dimensional simplex and contains the six vertices of the simplex. We calculate the value of the Bayes

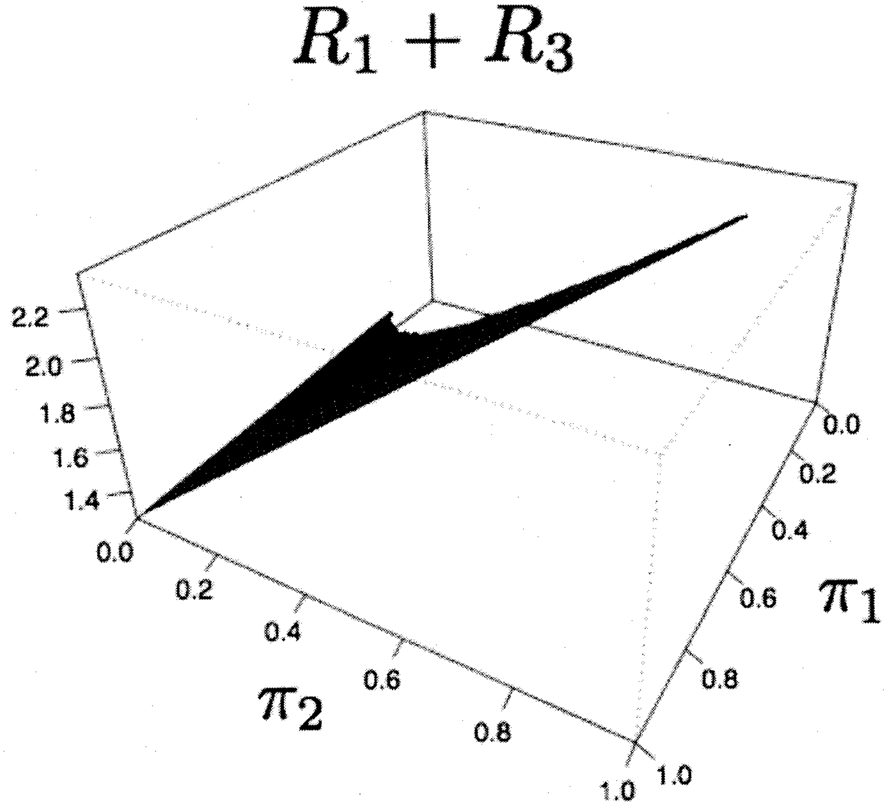


Figure 7.4: The terms $R_1 + R_3$ for the single-path model with hyperparameters $\bar{\mu}_1 = 19$, $\bar{\mu}_2 = 1$, $\bar{\sigma}_1^2 = 1$ and $\bar{\sigma}_2^2 = 2$ and variances $\sigma_1^2 = 1.5$ and $\sigma_2^2 = 1$.

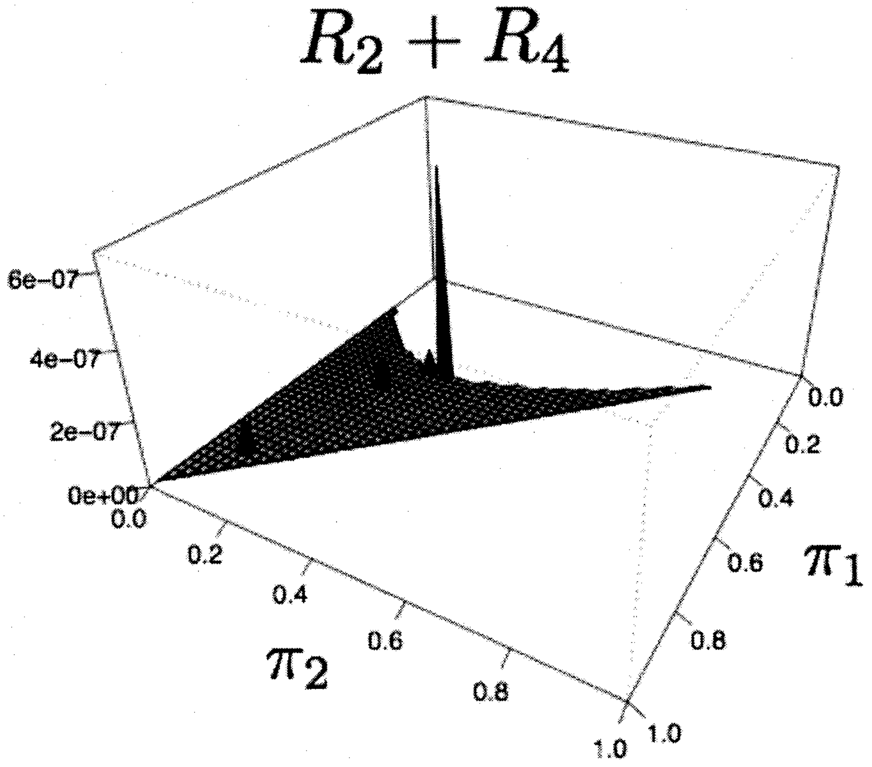


Figure 7.5: The terms $R_2 + R_4$ for the single-path model with hyperparameters $\bar{\mu}_1 = 19$, $\bar{\mu}_2 = 1$, $\bar{\sigma}_1^2 = 1$ and $\bar{\sigma}_2^2 = 2$ and variances $\sigma_1^2 = 1.5$ and $\sigma_2^2 = 1$. Note that $R_2 + R_4$ is essentially zero for these parameter values.

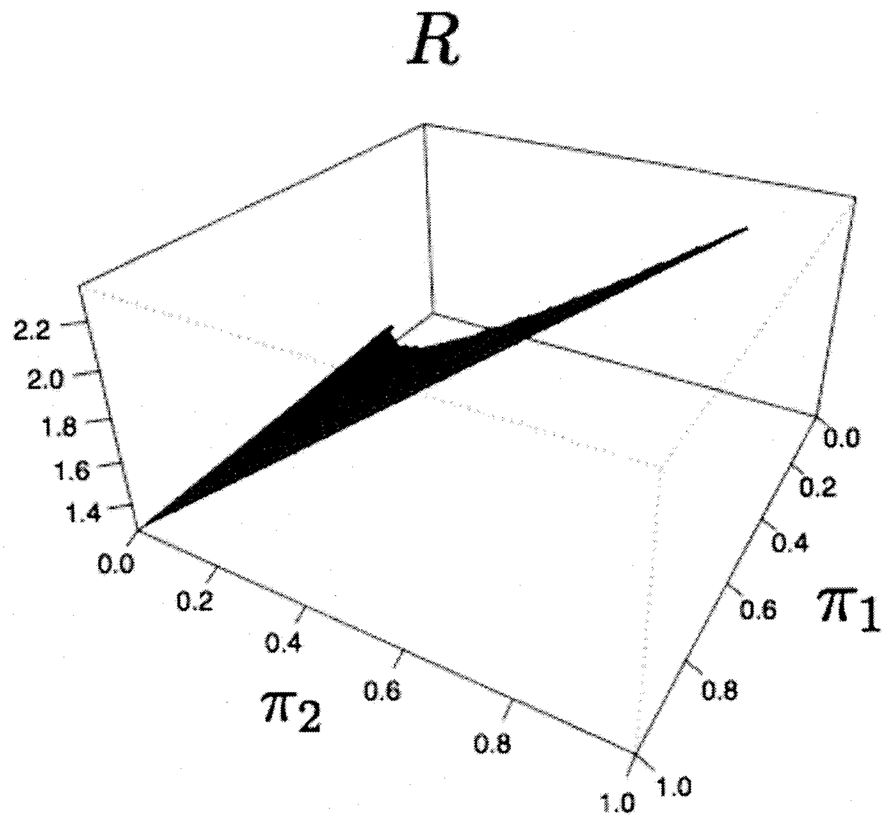


Figure 7.6: The Bayes risk based on squared error loss, R, for the single-path model with hyperparameters $\bar{\mu}_1 = 19$, $\bar{\mu}_2 = 1$, $\bar{\sigma}_1^2 = 1$ and $\bar{\sigma}_2^2 = 2$ and variances $\sigma_1^2 = 1.5$ and $\sigma_2^2 = 1$.

risk for the designs $\{u_0, u_1, u_2, u_3, u_4, u_5\}$ *in the set* V *which maps to the vertex set* $G_f(V)$. *By Theorem 6.4 we know that the optimal design is one of the designs in* V. *For comparison, we also compute the Bayes risk for the design where the points are equally spaced; we denote this design* u_{dist}. *The design with* π *having equal components is denoted by* u_{prob}.

Again we use normally distributed data and conjugate normal prior distributions for the before-and-after-change means. The truncated normal prior distribution of Figure 5.1 is used for the changepoint prior distribution. The hyperparameters are $\bar{\mu}_1 = 5$, $\bar{\mu}_2 = 4$, $\bar{\sigma}_1^2 = 1.5$ *and* $\bar{\sigma}_2^2 = 1.5$. *The variances of the model are* $\sigma_1^2 = 2$ *and* $\sigma_2^2 = 2$. *The interval* $[0, T]$ *has* T *equal to* 10. *The minimum distance* d *between design points equals* 0.5. *The results are shown in Table 7.3.*

Design	R_1	R_3	$R_1 + R_3$	$R_2 + R_4$	R
u_0	0.326	1.449	1.775	0.065	1.840
u_1	0.382	0.845	1.227	0.024	1.251
u_2	0.465	0.599	1.064	0.012	1.076
u_3	0.599	0.465	1.064	0.014	1.077
u_4	0.845	0.382	1.227	0.024	1.251
u_5	1.449	0.326	1.775	0.064	1.839
u_{dist}	0.589	0.589	1.178	0.263	1.442
u_{prob}	0.685	0.685	1.370	0.453	1.823

Table 7.3: Values from the numerical simulation for the designs in the set V and the designs u_{dist} and u_{prob}.

As expected, the R_1 *term is minimized by the design* u_0 *placing all observations towards* 0. *Likewise, the* R_3 *term is minimized by the design* u_5 *placing all observations towards* T. *Since the before-and-after-change hyperparametric and model variances are equal we see a symmetry in our results. The Bayes risks at* u_0 *and* u_5 *are equal,* *as are they at u_1 and u_4, and at u_2 and u_3. Any difference is due to numerical error. In particular, we see that with the odd number of measurements both u_2 and u_3 are optimal designs. For completeness we provide results for u_{dist}, where all points are the same distance apart, and for u_{prob}, where all points are an equal probability apart. We see that u_{prob} has a fairly high value for the Bayes risk, a value almost equal to the highest value obtained by u_0 and u_5. The design u_{dist}, although not having as high a value as u_{prob}, still has a fairly high value for the Bayes risk.*

7.2 The Common Changepoint Multi-Path Problem

We consider now the multi-path changepoint problem where all subjects have a common changepoint. As mentioned before, this assumption, although difficult to justify, may, in some cases, provide an approximation to the case for which the changepoints differ across subjects. We introduce random effects to allow subjects to have their own before-and-after-change means. The subjects' before-and-after-change means are normally distributed about hierarchical before-and-after-change means, respectively. We use the same design for all subjects. Our goal is to consider designs which provide the "best" estimates for the before-and-after-change hierarchical means. Again, the optimal design is found by minimizing the Bayes Risk based on squared error loss. As we will see, when the multi-path data are collapsed by averaging across subjects at each design point, the problem becomes a single-path changepoint problem with fixed correlation that does not depend on the design. Hence, all our single-path results from Chapter 6 apply to this common changepoint multi-path problem.

7.2.1 The Model

We consider m subjects and take n measurements on each subject. We assume the observations on each subject are conditionally independent given their before-and-after-change random effect means. The common design used on all subjects is again denoted by $x = (x_1, \ldots, x_n)$. Let i index the m subjects (i.e. $i = 1, \ldots, m$) and let j index the n measurements (i.e. $i = 1, \ldots, n$). The jth observation on the ith subject is denoted by y_{ij}. With $\tau_x = k$ we combine all observations on all subjects into the column vector y as follows.

$$y = (y_{11}, \ldots, y_{1k}, \ldots, y_{m1}, \ldots, y_{mk}, y_{1,k+1}, \ldots, y_{1n}, \ldots, y_{m,k+1}, \ldots, y_{mn})'. \qquad (7.21)$$

Let $y = ({y^1}', {y^2}')'$ where y^1 is the column vector of observations taken before the change

$$y^1 = (y_{11}, \ldots, y_{1k}, \ldots, y_{m1}, \ldots, y_{mk})' \qquad (7.22)$$

and y^2 is the vector of observations taken after the change

$$y^2 = (y_{1,k+1}, \ldots, y_{1n}, \ldots, y_{m,k+1}, \ldots, y_{mn})'. \qquad (7.23)$$

Let μ_{1i} and μ_{2i} denote the before-and-after-change means for subject i respectively. For simplicity of notation we combine the random effect before-and-after-change means for all subjects into vectors denoted by $m_1 = (\mu_{11}, ..., \mu_{1m})$ and $m_2 = (\mu_{21}, ..., \mu_{2m})$. The before-and-after hierarchical means are denoted by $\bar{\mu}_1$ and $\bar{\mu}_2$.

The hierarchical structure of our model is as follows. Consider the ith subject; if the observation j is taken before the changepoint we have:

$$\begin{aligned} y_{ij}|\mu_{1i} &\sim N(\mu_{1i}, \sigma_1^2) \\ \mu_{1i}|\bar{\mu}_1 &\sim N(\bar{\mu}_1, \bar{\sigma}_1^2) \qquad (7.24) \\ \bar{\mu}_1 &\sim N(\mu_1^*, \sigma_1^{2*}). \end{aligned}$$

Otherwise, if the observation j on the ith subject is taken after the changepoint we have:

$$\begin{aligned} y_{ij}|\mu_{2i} &\sim N(\mu_{2i}, \sigma_2^2) \\ \mu_{2i}|\bar{\mu}_2 &\sim N(\bar{\mu}_2, \bar{\sigma}_2^2) \\ \bar{\mu}_2 &\sim N(\mu_2^*, \sigma_2^{2*}). \end{aligned} \tag{7.25}$$

We denote the common changepoint for all subjects by τ and we represent the prior distribution of τ by $f(\tau)$. Again, the design measure is the probability mass function of the discrete random variable τ_x.

We assume that, given the random effect subject means m_1 and m_2, the data y are conditionally independent of the population means $\bar{\mu}_1$ and $\bar{\mu}_2$. Therefore, we have

$$f(y|m_1, m_2, \bar{\mu}_1, \bar{\mu}_2, \tau_x = k) = f(y|m_1, m_2, \tau_x = k). \tag{7.26}$$

Also, given each subject's before-and-after change means and the changepoint, all observations are independent. Hence, the likelihood is

$$f(y|m_1, m_2, \tau_x = k) = \prod_{i=1}^{m} \left(\prod_{j \leq k} f(y_{ij}|\mu_{1i}) \prod_{i > k} f(y_{ij}|\mu_{2i}) \right). \tag{7.27}$$

Recall from (7.24) and (7.25) in our model the distribution $f(y_{ij}|\mu_{1i})$ is $N(\mu_{1i}, \sigma_1^2)$ and the distribution $f(y_{ij}|\mu_{2i})$ is $N(\mu_{2i}, \sigma_2^2)$.

The parameters for the before-change random effect means are conditionally independent given the hierarchical before-change mean. Hence,

$$f(m_1|\bar{\mu}_1) = \prod_{i=1}^{m} f(\mu_{1i}|\bar{\mu}_1). \tag{7.28}$$

Similarly,

$$f(m_2|\bar{\mu}_2) = \prod_{i=1}^{m} f(\mu_{2i}|\bar{\mu}_2). \tag{7.29}$$

Since we ultimately wish to make inference about the hierarchical means $\bar{\mu}_1$ and $\bar{\mu}_2$, we begin by integrating out the random effect means m_1 and m_2 to find $f(y|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k)$.

Notice that without conditioning on the subject means m_1 and m_2, the observations for each subject need not be independent. In fact, the covariance matrix Σ_1 of observations taken before the changepoint y^1 has diagonal entries $\bar{\sigma}_1^2 + \sigma_1^2$ and off-diagonal entries $\bar{\sigma}_1^2$. Likewise, the covariance matrix Σ_2 for y^2 has diagonal entries $\bar{\sigma}_2^2 + \sigma_2^2$ and off-diagonal entries $\bar{\sigma}_2^2$. As shown in Section B.1 of the Appendix, using e to represent a column vector of 1's, we have, with an abuse of notation,

$$f(y|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k) = \prod_{i=1}^{m} N_k(\bar{\mu}_1 e, \Sigma_1) N_{(n-k)}(\bar{\mu}_2 e, \Sigma_2). \tag{7.30}$$

Using y^1, y^2 and the Kronecker product $\otimes$, we can also express the density as

$$\begin{aligned} &f(y^1, y^2|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k) \\ &= N_{mk}(e_m \otimes \bar{\mu}_1 e, I_m \otimes \Sigma_1) N_{m(n-k)}(e_m \otimes \bar{\mu}_2 e, I_m \otimes \Sigma_2). \end{aligned} \tag{7.31}$$

Next, we show that by taking the average of our data over the m subjects at each design point, we obtain a single-path changepoint problem. That is, the vector of sample means of the m observations at each design point is a sufficient statistic for $\bar{\mu}_1$ and $\bar{\mu}_2$. We denote the column vector of sample means as $\bar{y} = (\bar{y}_1, \dots, \bar{y}_n)'$, where $\bar{y}_j = \frac{1}{m}\sum_{i=1}^m y_{ij}$ is the average data collected at design point x_j. As calculated in Section B.2 of the Appendix,

$$f(\bar{y}|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k) = N_k(\bar{\mu}_1 e, \bar{\Sigma}_1) N_{(n-k)}(\bar{\mu}_2 e, \bar{\Sigma}_2). \tag{7.32}$$

Theorem 7.3. *The common changepoint multi-path model is a single-path changepoint model with before-and-after-change means $\bar{\mu}_1$ and $\bar{\mu}_2$ when the sequence $\bar{y}$ is considered.*

Proof. The result follows immediately from the form of $f(\bar{y}|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k)$ in expression (7.32). The prior distributions for $\bar{\mu}_1$ and $\bar{\mu}_2$ are $N(\mu_1^*, \sigma_1^{*2})$ and $N(\mu_2^*, \sigma_2^{*2})$. □

Define the column vectors

$$\bar{y}_1 = (\bar{y}_1, \dots, \bar{y}_k)' \tag{7.33}$$

and

$$\bar{y}^2 = (\bar{y}_{k+1}, \ldots, \bar{y}_n)'. \tag{7.34}$$

From the form of expression (7.32), given the parameters $\bar{\mu}_1$, $\bar{\mu}_2$, and $\tau_x = k$, the averages $\bar{y}^1$ taken before the change are independent of the averages $\bar{y}^2$ taken after the change. In addition, the observations $\bar{y}^1$ have a fixed correlation between them, not depending on the distances between the design points $(x_1, \ldots, x_k)$. Similarly, $\bar{y}^2$ also have a fixed correlation between them not depending on the distances between the design points $(x_{k+1}, \ldots, x_n)$. Since the common changepoint multi-path problem is a single-path changepoint problem with a correlation not depending on the design, all our optimal design results from Chapter 6 apply. Next we consider the Bayes risk based on squared error loss for estimation of the before-and-after-change hierarchical means.

7.2.2 The Bayes Risk Based on Squared Error Loss

From Theorem 7.3 we have a single-path changepoint problem for the sequence of averages $\bar{y}$. Following the same steps as in Chapter 6 we decompose the Bayes risk based on squared error loss into the four terms R_1, R_2, R_3 and R_4:

$$\begin{aligned}
R = & \int E_{\tau_x=k|\bar{y}}(Var(\bar{\mu}_1|\bar{y}, \tau_x = k))f(\bar{y})d\bar{y} \\
& + \int Var_{\tau_x=k|\bar{y}}(E(\bar{\mu}_1|\bar{y}, \tau_x = k))f(\bar{y})d\bar{y} \\
& + \int E_{\tau_x=k|\bar{y}}(Var(\bar{\mu}_2|\bar{y}, \tau_x = k))f(\bar{y})d\bar{y} \\
& + \int Var_{\tau_x=k|\bar{y}}(E(\bar{\mu}_2|\bar{y}, \tau_x = k))f(\bar{y})d\bar{y}.
\end{aligned} \tag{7.35}$$

The terms R_1 and R_3 can be integrated, while the terms R_2 and R_4 lead to a non-linear integral in $\bar{y}$.

To calculate the terms of the Bayes risk (7.35), we need the posterior expectations and variances of $\bar{\mu}_1$ and $\bar{\mu}_2$ given $\bar{y}$ and $\tau_x = k$ and the densities $f(\tau_x = k|\bar{y})$ and $f(\bar{y})$. We calculate the posterior expectations and variances of $\bar{\mu}_1$ and $\bar{\mu}_2$ given $\bar{y}$ and $\tau_x = k$ in Appendix B.3. The densities $f(\tau_x = k|\bar{y})$ and $f(\bar{y})$ are easily obtained using equations (7.36) and (7.37) once the density $f(\bar{y}|\tau_x = k)$ is known. We obtain the density $f(\bar{y}|\tau_x = k)$ in Appendix B.4. Hence we have

$$f(\bar{y}) = \sum_{k=0}^{n} f(\bar{y}|\tau_x = k)\pi_k \tag{7.36}$$

and

$$f(\tau_x = k|\bar{y}) = \frac{f(\bar{y}|\tau_x = k)\pi_k}{f(\bar{y})}. \tag{7.37}$$

7.2.3 Simulations

We present two examples of the common changepoint multi-path problem. We use two design points so that the Bayes risk can easily be plotted, and we focus on how the magnitude of the risk changes as m, the number of subjects, increases.

Example 7.5. Two Subjects

In Table 7.4 and Figures 7.7, 7.8 and 7.9 we present numerical simulations of the Bayes risk for the common changepoint multi-path model with two subjects, hyperparameters $\mu_1^* = 4.5$, $\mu_2^* = 4$, $\sigma_1^{*2} = 3$, $\sigma_2^{*2} = 2$, $\bar{\sigma}_1^2 = 2.5$ and $\bar{\sigma}_2^2 = 3$ and model variances $\sigma_1^2 = 3$ and $\sigma_2^2 = 3$. The truncated log-concave normal prior distribution of Figure 5.1 for the changepoint prior distribution. The minimum distance d is 2 and T is 10.

From Table 7.4 we see that the Bayes risk is minimized at the design u_1. This table also shows that R_1 is minimized by the design u_0, while R_3 is minimized by the design u_2. The sum $R_1 + R_3$ is minimized at the design u_1. These are the same features that we saw in Examples 7.2 and 7.3.

Figures 7.7, 7.8, and 7.9 plot $R_1 + R_3$, $R_2 + R_4$, and R, respectively. We see the linear structure of $R_1 + R_3$ in Figure 7.7 and the concave form of $R_2 + R_4$ in

7.8. Figure 7.9 displays the complete Bayes risk R for these model variances and hyperparameters.

Design	R_1	R_3	$R_1 + R_3$	$R_2 + R_4$	R
u_0	1.251	1.932	3.183	0.140	3.323
u_1	1.467	1.971	2.664	0.0411	2.704
u_2	2.776	1.079	3.855	0.221	4.076

Table 7.4: Values from the numerical simulation for the designs in the set V when there are two subjects. Due to the numerical simulation of $R_2 + R_4$ and of region $G_f(\mathbb{X}^n)$, these values are approximate. For instance, clearly the $R_2 + R_4$ term should be 0 for u_1.

Example 7.6. Three Subjects

We present a similar example to Example 7.5, except that we now have three subjects. In Table 7.5 and Figures 7.10, 7.11 and 7.9, we present numerical simulations of the Bayes risk for this problem.

From Table 7.5 we make the same observations as before: the Bayes risk is minimized by the design u_1. Once more, R_1 is minimized at the design u_0 and R_3 is minimized at the u_2 design. The sum $R_1 + R_3$ is minimized at the design u_1. These are same same features that we saw in Examples 7.2, 7.3 and 7.5.

Figures 7.10, 7.11 and 7.12 plot $R_1 + R_3$, $R_2 + R_4$ and R respectively. We see the linear forms of $R_1 + R_3$ in Figure 7.10 and the concave form of $R_2 + R_4$ in 7.11. In Figure 7.12 we see the complete Bayes risk R for these model variances and hyperparameters.

Evident from Table 7.5 and Figures 7.10, 7.11 and 7.12 is that adding a third subject reduces the magnitude of the risk.

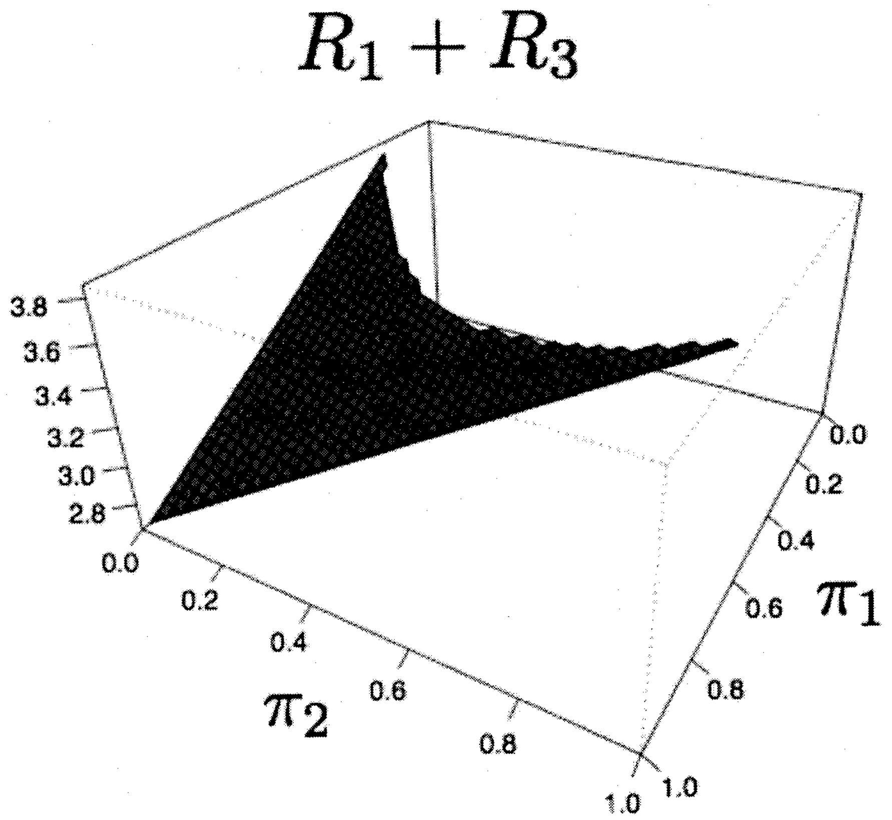


Figure 7.7: The term $R_1 + R_3$ for the common changepoint multi-path model with two subjects and hyperparameters $\mu_1^* = 4.5$, $\mu_2^* = 4$, $\sigma_1^{*2} = 3$, $\sigma_2^{*2} = 2$, $\bar{\sigma}_1^2 = 2.5$ and $\bar{\sigma}_2^2 = 3$. The variances of the model are $\sigma_1^2 = 3$ and $\sigma_2^2 = 3$.

Design	R_1	R_3	$R_1 + R_3$	$R_2 + R_4$	R
u_0	0.979	1.915	2.894	0.114	3.008
u_1	1.759	0.997	2.173	0.031	2.204
u_2	2.734	0.878	3.612	0.129	3.741

Table 7.5: Values from the numerical simulation for the designs in the set V for the common changepoint multi-path subject when there are three subjects. Due to the numerical simulation of $R_2 + R_4$, and the region $G_f(\mathbb{X}^n)$, these values are approximate. For instance, clearly $R_2 + R_4$ should be 0 for u_1.

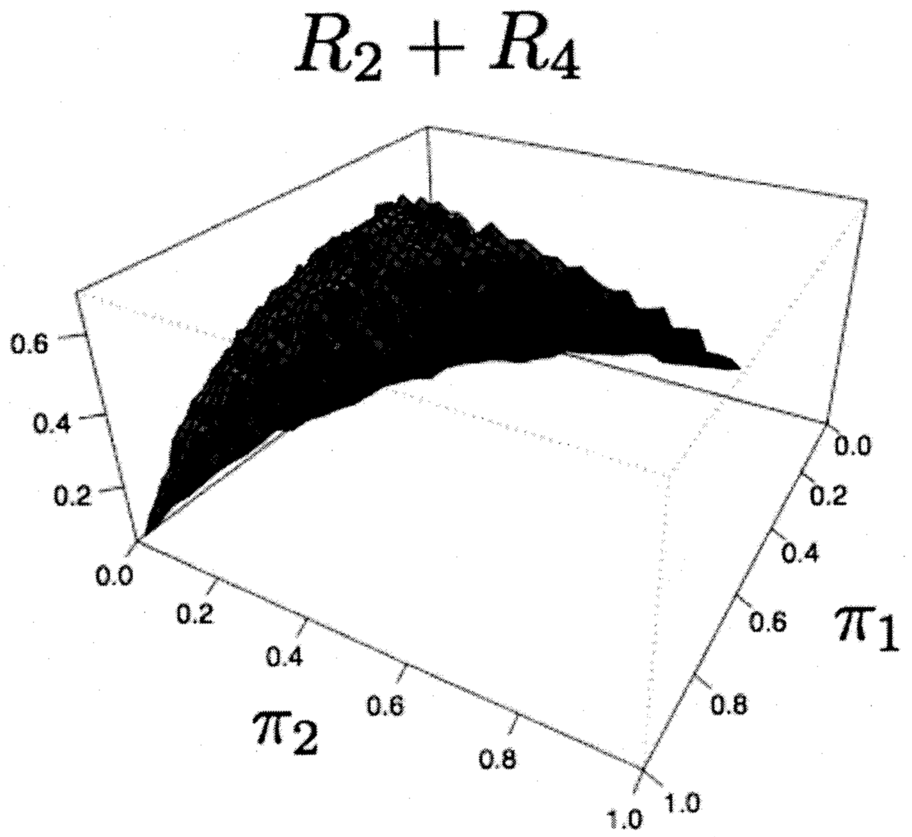


Figure 7.8: The term $R_2 + R_4$ for the common changepoint multi-path model with two subjects and hyperparameters $\mu_1^* = 4.5$, $\mu_2^* = 4$, $\sigma_1^{*2} = 3$, $\sigma_2^{*2} = 2$, $\bar{\sigma}_1^2 = 2.5$ and $\bar{\sigma}_2^2 = 3$. The variances of the model are $\sigma_1^2 = 3$ and $\sigma_2^2 = 3$.

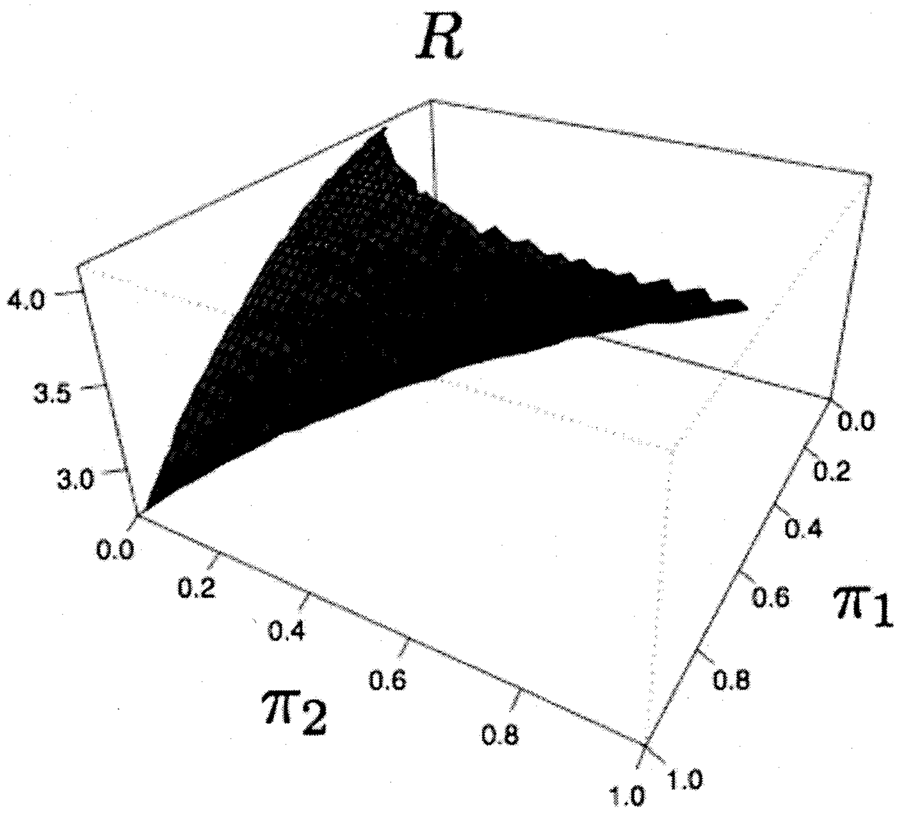


Figure 7.9: The Bayes risk based on squared error loss R for the common changepoint multi-path model with two subjects and hyperparameters $\mu_1^* = 4.5$, $\mu_2^* = 4$, $\sigma_1^{*2} = 3$, $\sigma_2^{*2} = 2$, $\bar{\sigma}_1^2 = 2.5$ and $\bar{\sigma}_2^2 = 3$. The variances of the model are $\sigma_1^2 = 3$ and $\sigma_2^2 = 3$.

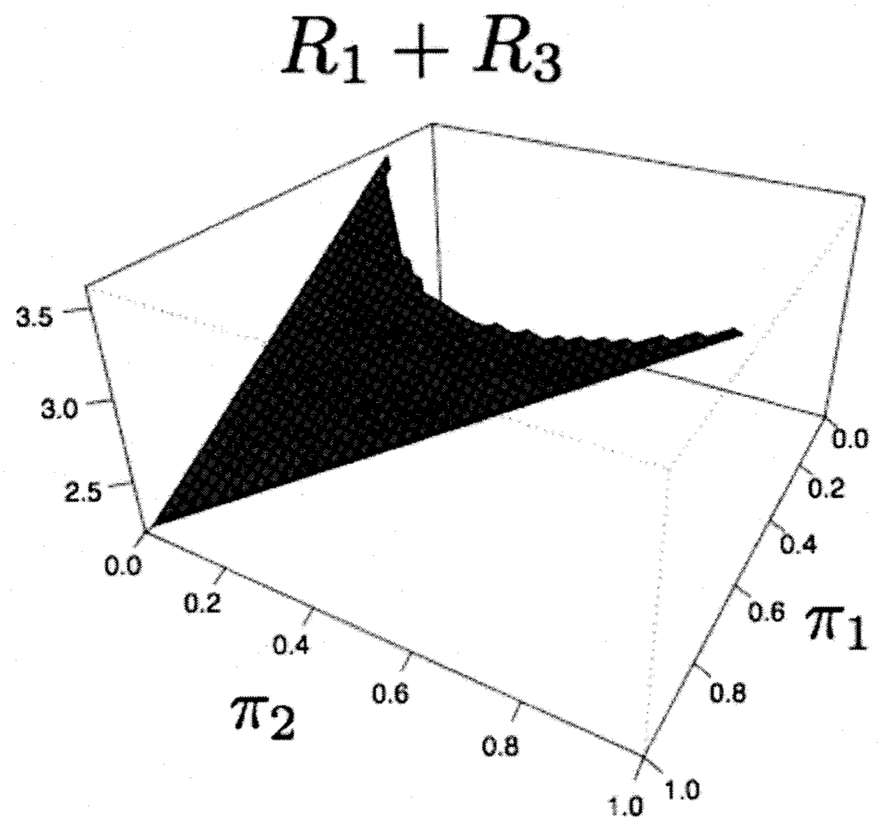


Figure 7.10: The term $R_1 + R_3$ for the common changepoint multi-path model with three subjects and hyperparameters $\mu_1^* = 4.5$, $\mu_2^* = 4$, $\sigma_1^{*2} = 3$, $\sigma_2^{*2} = 2$, $\bar{\sigma}_1^2 = 2.5$ and $\bar{\sigma}_2^2 = 3$. The variances of the model are $\sigma_1^2 = 3$ and $\sigma_2^2 = 3$.

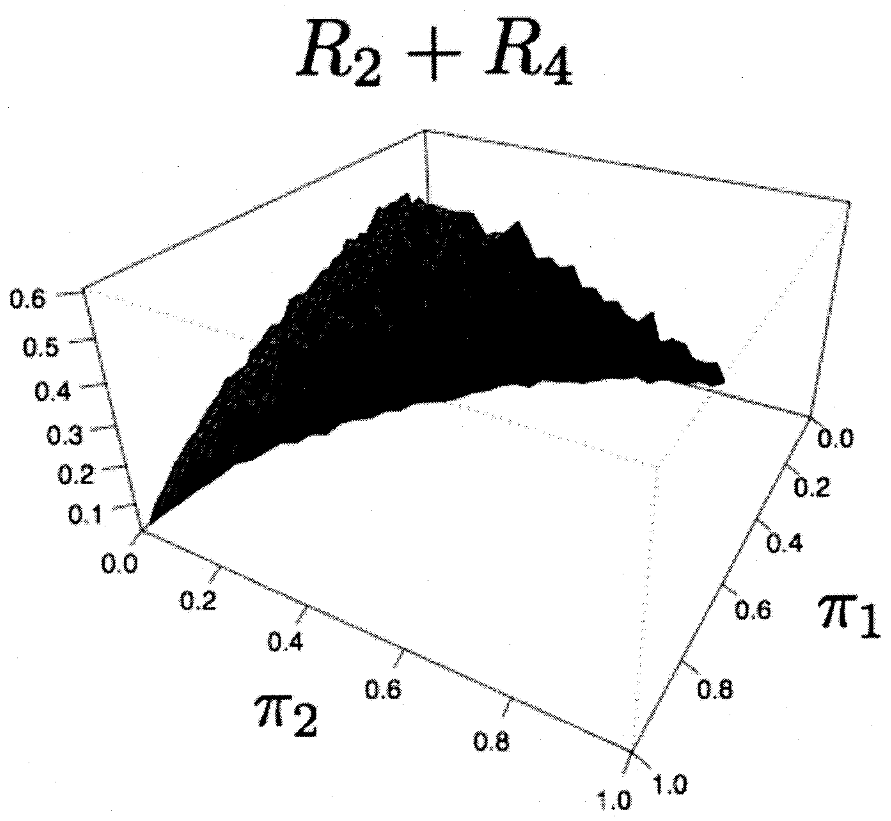


Figure 7.11: The term $R_2 + R_4$ for the common changepoint multi-path model with three subjects and hyperparameters $\mu_1^* = 4.5$, $\mu_2^* = 4$, $\sigma_1^{*2} = 3$, $\sigma_2^{*2} = 2$, $\bar{\sigma}_1^2 = 2.5$ and $\bar{\sigma}_2^2 = 3$. The variances of the model are $\sigma_1^2 = 3$ and $\sigma_2^2 = 3$.

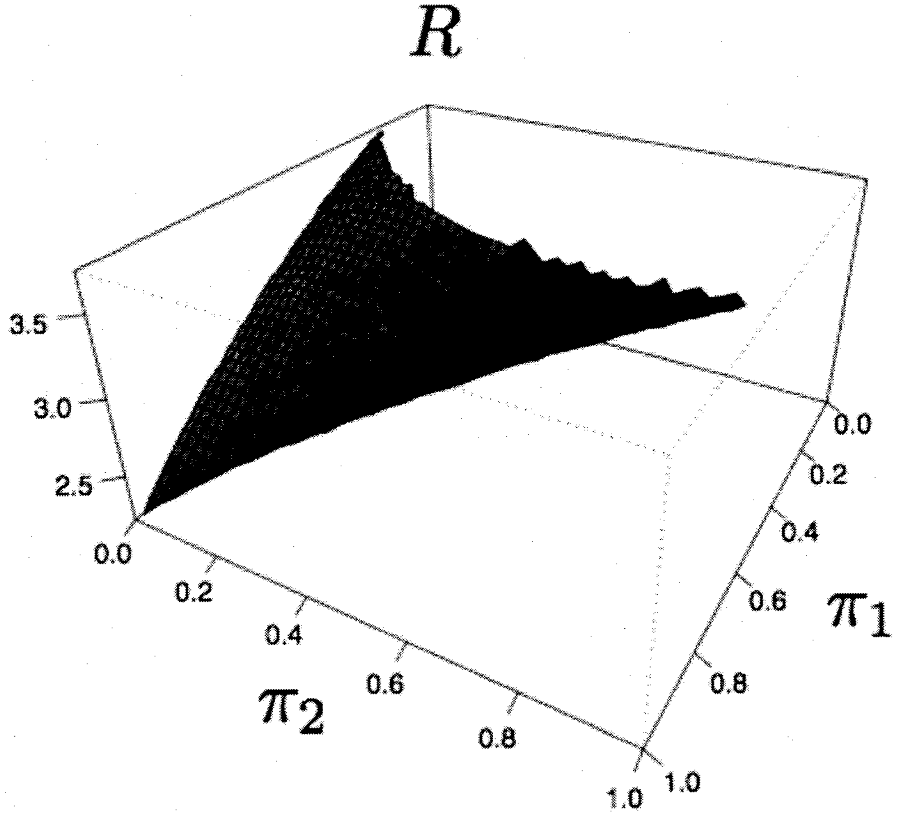


Figure 7.12: The Bayes risk based on squared error loss, R, for the common changepoint multi-path model with three subjects and hyperparameters $\mu_1^* = 4.5$, $\mu_2^* = 4$, $\sigma_1^{*2} = 3$, $\sigma_2^{*2} = 2$, $\bar{\sigma}_1^2 = 2.5$ and $\bar{\sigma}_2^2 = 3$. The variances of the model are $\sigma_1^2 = 3$ and $\sigma_2^2 = 3$.

Chapter 8

Optimal Design for the Multiple Changepoints Multi-Path Problem

We consider the multiple changepoints multi-path problem. This problem is exactly like the common changepoint multi-path problem of Section 7.2, except that each subject has his or her own changepoint. For instance, in the motivational blood pressure example in Chapter 1 of this thesis, the treatment is not likely to take effect at exactly the same time in all subjects.

As in the common changepoint multi-path problem, we have hierarchical before-and-after-change means. Subject random effect before-change means are normally distributed about the hierarchical before-change mean. Likewise, subject random effect after-change means are normally distributed about the hierarchical after-change mean. Conjugate normal prior distributions are assigned to the before-and-after-change hierarchical means. The random variables for the subject specific changepoints are independent and identically distributed.

The work in this chapter is preliminary and, unlike previous chapters, we do not have complete optimal design results. We include this chapter because it addresses optimal design for the multi-path situation of Joseph et al. (1996), which motivated

Example 1.1. We use the squared error loss Bayes risk for the three optimal design problems in this chapter.

The first problem we consider is to find the optimal design when the goal is to estimate the proportion of subjects who do not undergo a change. For instance, in Joseph et al. (1996), a blood pressure treatment was administered to many people and the treatment had no effect on a proportion of the people. Interest may lie in estimating the proportion of the population unaffected by the treatment.

Similarly, we also consider optimal designs for estimating the proportion of subjects who undergo a change in a specific interval. Returning again to the blood pressure example, we might be interested in estimating the proportion of people affected by the treatment during a certain interval after it was administered. That is, if the treatment was administered at time t_1 we might wish to estimate the number of people affected in the interval $[t_1, t_2]$.

It is unclear whether or not the criterion functions for these two problems are concave functions of π. The complication is due to the facts that there are multiple changepoints and that measurements from different sequences can be correlated because of the subject before-and-after-change random effect means. We consider the two optimal design problems of estimating the proportion of subjects who do not change and the proportion of subjects who change in a specific subinterval in Section 8.2 after we introduce the multiple changepoints multi-path model in Section 8.1.

In Section 8.3 we consider the third optimal design problem of estimating the before-and-after-change hierarchical means. We show that the Bayes risk is not always concave and hence the optimal design is not necessarily one which places observations at the ends of the interval.

As we will see, although it arises in a slightly more complicated fashion, the same design measure π is present in this multiple changepoints multi-path problem. We are taking repeated measurements on each subject and therefore must again keep design points a minimum distance d apart. We are thus still minimizing over $G_f(\mathbb{X}^n)$ in this multiple changepoints setting.

8.1 The Model

We consider m independent subjects and take n measurements on each subject. We assume that a common design is used on all subjects, and denote it by $x = (x_1, \ldots, x_n)$. As before, design points are taken at least a distance d apart and hence the set of allowable designs is $\mathbb{X}^n$. Let i index the m subjects (i.e. $i = 1, \ldots, m$) and let j index the n measurements (i.e. $j = 1, \ldots, n$). Therefore we denote the jth observation on the ith subject by y_{ij}.

We denote the before-and-after-change means for subject i by μ_{1i} and μ_{2i}. For simplicity of notation, we combine the before-and-after-change means for all subjects into vectors denoted by $m_1 = (\mu_{11}, ..., \mu_{1m})$ and $m_2 = (\mu_{21}, ..., \mu_{2m})$. We use the parameters $\bar{\mu}_1$ and $\bar{\mu}_2$ to denote the before-and-after population or hierarchical means.

The changepoint for subject i is denoted by τ_i and we use τ, here, to represent the row vector of all changepoints. Letting g denote the multivariate changepoint density, we assume that the changepoints of the m subjects are identically and independently distributed with joint density

$$g(\tau) = \prod_{i=1}^{m} f(\tau_i). \tag{8.1}$$

To obtain the design measure for this problem, we combine the multivariate changepoint random vector τ with the design x to create a new discrete multivariate random vector τ_x. Taking x_0 to be 0 and x_{n+1} to be T, the probability mass function for τ_{xi}, the ith component of τ_x, is given below:

$$\pi_k \overset{\text{def}}{=} P(\tau_{xi} = k_i) = \int_{x_{k_i}}^{x_{k_i}+1} f. \tag{8.2}$$

Hence density of τ_x is

$$P(\tau_x = k) = P\big((\tau_{x1}, \ldots, \tau_{xm}) = (k_1, \ldots, k_m)\big) = \pi_{k_1} \cdots \pi_{k_m}.$$

As before we combine the design measure and the changepoint vector τ, and re-express our design criterion functions in terms of the new multivariate discrete random vector τ_x rather than τ and x separately. Hence, we again recast the problem of minimizing the Bayes risk over the design space $\mathbb{X}^n$ into one of minimizing the risk over the design measure space $G_f(\mathbb{X}^n)$.

As in Section 7.2.1 we create the column vectors y, containing all observations, y^1, containing observations taken before the change, and y^2, containing observations taken after the change. Since each subject has his or her own changepoint, y, y^1 and y^2 have the following forms:

$$y = (y_{11}, \ldots, y_{1k_1}, \ldots, y_{m1}, \ldots, y_{mk_m}, y_{1,k_1+1}, \ldots, y_{1n}, \ldots, y_{m,k_m+1}, \ldots, y_{mn})', \quad (8.3)$$

$$y^1 = (y_{11}, \ldots, y_{1k}, \ldots, y_{m1}, \ldots, y_{mk})', \quad (8.4)$$

and

$$y^2 = (y_{1,k+1}, \ldots, y_{1n}, \ldots, y_{m,k+1}, \ldots, y_{mn})'. \quad (8.5)$$

The hierarchical structure of our model is as follows: consider the ith subject. If, jth observation on subject i is taken before the changepoint τ_i we have:

$$\begin{aligned} y_{ij}|\mu_{1i} &\sim N(\mu_{1i}, \sigma_1^2) \\ \mu_{1i}|\bar{\mu}_1 &\sim N(\bar{\mu}_1, \bar{\sigma}_1^2) \\ \bar{\mu}_1 &\sim N(\mu_1^*, \sigma_1^{2*}). \end{aligned} \quad (8.6)$$

Otherwise, if the jth observation on the ith subject is taken after the changepoint τ_i, we have:

$$\begin{aligned} y_{ij}|\mu_{2i} &\sim N(\mu_{2i}, \sigma_2^2) \\ \mu_{2i}|\bar{\mu}_2 &\sim N(\bar{\mu}_2, \bar{\sigma}_2^2) \\ \bar{\mu}_2 &\sim N(\mu_2^*, \sigma_2^{2*}). \end{aligned} \quad (8.7)$$

As with the common changepoint multi-path problem, we assume that given the subject mean vectors m_1 and m_2, the data y are conditionally independent of the hierarchical means $\bar{\mu}_1$ and $\bar{\mu}_2$. That is, once we know the random effect means, the hierarchical means provide no extra information. Therefore,

$$f(y|m_1, m_2, \bar{\mu}_1, \bar{\mu}_2, \tau_x = k) = f(y|m_1, m_2, \tau_x = k). \tag{8.8}$$

Assuming that, given the before-and-after-change means, all observations are independent, we have

$$f(y|m_1, m_2, \tau_x = k) = \prod_{i=1}^{m} \left(\prod_{j \le k_i} f(y_{ij}|\mu_{1i}) \prod_{j > k_i} f(y_{ij}|\mu_{2i}) \right) \tag{8.9}$$

Further, assuming that the subject specific means are conditionally independent, given $\bar{\mu}_1$ and $\bar{\mu}_2$, respectively, we have

$$f(m_1|\bar{\mu}_1) = \prod_{i=1}^{m} f(\mu_{1i}|\bar{\mu}_1) \tag{8.10}$$

and

$$f(m_2|\bar{\mu}_2) = \prod_{i=1}^{m} f(\mu_{2i}|\bar{\mu}_2). \tag{8.11}$$

Finally, we suppose that, $\bar{\mu}_1 \sim N(\mu_1^*, \sigma_1^{*2})$ and $\bar{\mu}_2 \sim N(\mu_2^*, \sigma_2^{*2})$.

8.2 Estimation of Proportions

Consider the design problems for estimating either the proportion of people who do not undergo a change or the proportion of people who change in the subinterval $[t_1, t_2]$. We wish to find optimal designs for estimating

$$\frac{\sum_{i=1}^{m} I_{\tau_{xi}=n}}{m} \tag{8.12}$$

while fixing x_n at T, and for estimating

$$\frac{\sum_{i=1}^{m} I_{\tau_{xi}=q}}{m} \tag{8.13}$$

while fixing x_q at t_1 and x_{q+1} at t_2. The design criterion functions based on the squared error loss Bayes risk are respectively, fixing x_n at T,

$$\int Var\left(\frac{\sum_{i=1}^m I_{\tau_{xi}=n}}{m}\middle| y\right) f(y)dy \tag{8.14}$$

and fixing x_q at t_1 and x_{q+1} at t_2,

$$\int Var\left(\frac{\sum_{i=1}^m I_{\tau_{xi}=q}}{m}\middle| y\right) f(y)dy. \tag{8.15}$$

Owing to the hierarchical before-and-after-change means and the random effects, although the τ_{xi} are independent, they are not conditionally independent, given the data y. (See Appendix C.3, where the density $f(y|\tau_x = k)$ is presented). As a result, the expressions (8.14) and (8.15) are quite complicated functions of $(\pi_0, \ldots, \pi_{n-1})$ and $(\pi_0, \ldots, \pi_{q-1}, \hat{\pi}, \pi_{q+1}, \ldots, \pi_n)$ respectively. Furthermore, it seems likely that they are not concave functions of π. If so, it could be that the optimal design is not one of the designs in the set V.

8.3 Estimation of the Hierarchical Before-and-After-Change Means

As in the common changepoint multi-path problem, we use the Bayes risk based on squared error loss to estimate the before-and-after-change hierarchical means. We decompose the Bayes risk into four parts (R_1, R_2, R_3, R_4) reproduced below for easy reference. Recall that the difference between the current setting and the common changepoint multi-path problem is that τ_x is now multivariate.

$$\begin{aligned} R = & \int E_{\tau_x=k|y}(Var(\bar{\mu}_1|y, \tau_x = k))f(y)dy \\ & + \int Var_{\tau_x=k|y}(E(\bar{\mu}_1|y, \tau_x = k))f(y)dy \\ & + \int E_{\tau_x=k|y}(Var(\bar{\mu}_2|y, \tau_x = k))f(y)dy \\ & + \int Var_{\tau_x=k|y}(E(\bar{\mu}_2|y, \tau_x = k))f(y)dy \end{aligned} \tag{8.16}$$

It turns out that the optimal design is not always one of the designs placing the observations as far as possible at the ends of the observation interval. The Bayes risk for this problem is not always concave.

In Appendix C.3 we calculate $f(y|\tau_x = k)$. From $f(y|\tau_x = k)$ we can easily find $f(y)$ and $f(\tau_x = k|y)$, as shown below.

$$\begin{aligned} f(y) &= \sum_{(k_1,\dots,k_m)} f(y|\tau_x = k)p(\tau_x = k) \\ &= \sum_{(k_1,\dots,k_m)} f(y|\tau_x = k)\pi_{k_1}\cdots\pi_{k_m} \end{aligned} \tag{8.17}$$

$$\begin{aligned} f(\tau_x = k|y) &= \frac{f(y|\tau_x = k)p(\tau_x = k)}{f(y)} \\ &= \frac{f(y|\tau_x = k)\pi_{k_1}\cdots\pi_{k_m}}{f(y)} \end{aligned} \tag{8.18}$$

In Section C.2 of Appendix C we calculate $E(\bar{\mu}_1|y, \tau_x = k)$, $Var(\bar{\mu}_1|y, \tau_x = k)$, $E(\bar{\mu}_2|y, \tau_x = k)$ and $Var(\bar{\mu}_2|y, \tau_x = k)$.

By Fubini's Theorem, we find

$$\begin{aligned} R_1 &= \int E_{\tau_x=k|y}(Var(\bar{\mu}_1|y, \tau_x = k))f(y)dy \\ &= \sum_{(k_1,\dots,k_m)} E_{y|\tau_x=k}Var(\bar{\mu}_1|y, \tau_x = k)\pi_{k_1}\cdots\pi_{k_m}. \end{aligned} \tag{8.19}$$

The term R_3 is found similarly. Furthermore, we find for this multiple changepoints multi-path problem that R_1 and R_3 are no longer linear functions of π. In fact we find that they are convex functions of π, and that the curvature increases with the number m of subjects.

The terms R_2 and R_4 are very complicated functions of π and it is not clear whether they can be concave functions of π or not. Examples 8.1 and 8.2 provide situations where $R_2 + R_4$ are not concave.

8.4 Simulations

We consider two simple numerical examples with a single design point to investigate the Bayes risk based on squared error loss for estimating the before-and-after-change hierarchical means. In particular, we consider the concavity of the terms R_1, R_3, R_2 and R_4 and we strive to understand why the optimal design might not be one of the designs in the set V.

Example 8.1. One Design Point

We took the model variances to be $\sigma_1^2 = 2$ and $\sigma_2^2 = 1$ respectively, and the hyperparameters to be $\bar{\sigma}_1^2 = 2$, $\bar{\sigma}_2^2 = 1$, $\sigma_1^{*2} = 1$, $\sigma_2^{*2} = 1$, $\mu_1^* = 4$, and $\mu_2^* = 1$. The number of subjects was 3. As seen in Figures 8.1 (a) and (b) R_1 and R_3 are slightly convex. From Figure 8.1 (c) we see that $R_2 + R_4$ is certainly not concave. In this example, the shape of the Bayes risk most closely resembles $R_2 + R_4$. Note that even though the Bayes risk is not concave in this example, the optimal design will be one of the designs placing the design point at 0 or at T.

In the next example, we take parameter and hyperparameter values which demonstrate that the optimal design may not be to position a measurement at 0 or T. That is, the optimal design places the measurement somewhere in the interior of the interval $[0, T]$.

Example 8.2. One Design Point *Here we took the same model and hyperparametric variances as in Example 8.1. However we took $\mu_1^* = 20$ and $\mu_2^* = 1$. The number*

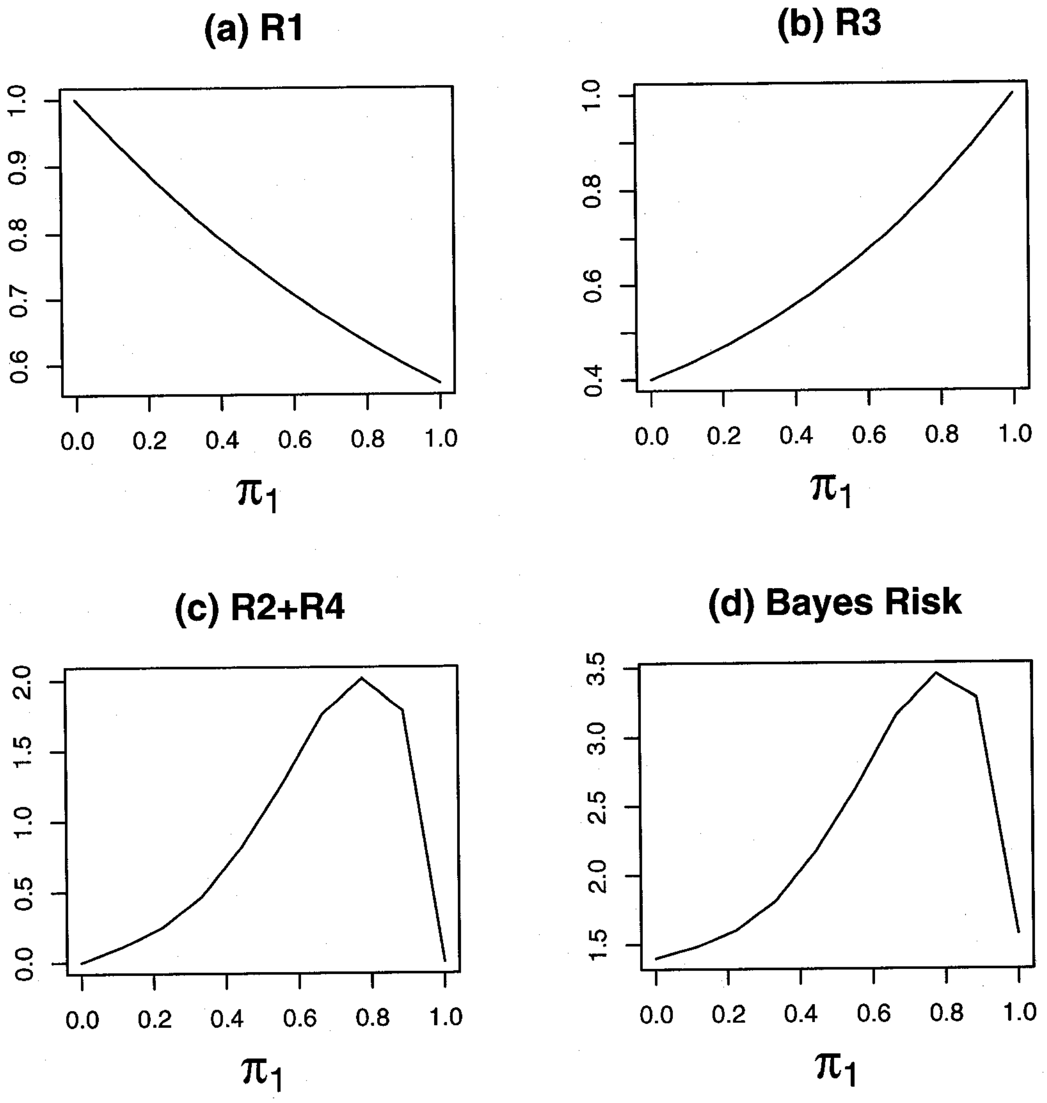


Figure 8.1: (a) R_1, (b) R_3, (c) R_2+R_4 and the (d) Bayes risk R based on squared error loss for the multiple changepoints multi-path model, with model variances $\sigma_1^2 = 2$ and $\sigma_2^2 = 1$ and hyperparameters $\bar{\sigma}_1^2 = 2$, $\bar{\sigma}_2^2 = 1$, $\sigma_1^{*2} = 1$, $\sigma_2^{*2} = 1$, $\mu_1^* = 4$, and $\mu_2^* = 1$.

of subjects is three. From our observation in Chapter 7 (that $R_1 + R_3$ dominates the Bayes risk when the hyperparametric before-and-after-change means are far apart compared to the variances), we expect the $R_1 + R_3$ to dominate in this example. As seen in Figures 8.2 (a) and (b), R_1 and R_3 are again slightly convex. From Figure 8.2 (c) we see that $R_2 + R_4$ is essentially zero. Figure 8.2 (d) shows that the Bayes risk is approximately $R_1 + R_3$. Here we have an example where the optimal design is to place the design point in the interior of $[0, T]$. The exact placement of the design point would depend on the prior distribution for the changepoint.

One might wonder if the optimal design that places a point in the interior of $[0, T]$, is just an artifact of having a single design point. Hence we present a third example below, with three design points.

Example 8.3. Three Design Points *Consider a multiple changepoints multi-path model with three design points and three subjects. We used the AMPL software to minimize $R_1 + R_3$ taking all model variances and hyperparametric variances to be one. Since R_1 and R_3 do not depend on the hyperparametric before-and-after means we assumed these means were far enough apart so that $R_2 + R_4$ was negligible. For this example the minimum was at $(\pi_0, \pi_1, \pi_2, \pi_3) = (0, 0.5, 0.5, 0)$. The exact placement of the points depends on the prior distribution for the changepoint, although it is clear that the middle design point will be well into the interior of $[0, T]$ for most prior distributions on the changepoint.*

Next, using a single design point in Example 8.4, we illustrate why the optimal design would place a design point in the interior of $[0, T]$.

Example 8.4. *Consider the three designs in Figure 8.3. In the first design (a) all three measurements will be taken from the "μ_1 distribution", and in the third design (c) all three measurements will be taken from the "μ_2 distribution". The second design (b) is a compromise because, since the individual changepoints occur at different*

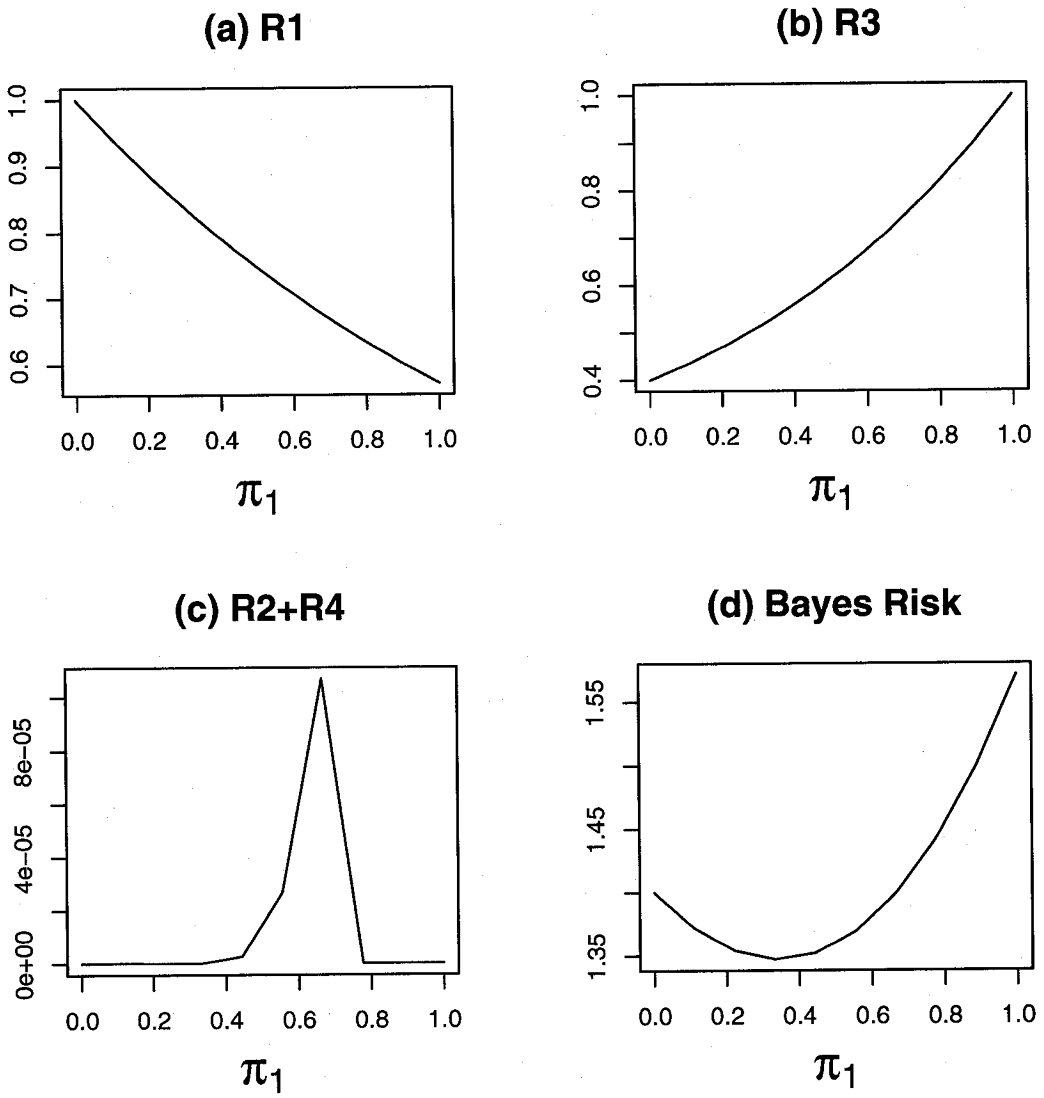


Figure 8.2: (a) R_1, (b) R_3, (c) $R_2 + R_4$ (d) Bayes risk, R, based on squared error loss for the multiple changepoints multi-path model, with model variances $\sigma_1^2 = 2$ and $\sigma_2^2 = 1$, and hyperparameters $\bar{\sigma}_1^2 = 2$, $\bar{\sigma}_2^2 = 1$, $\sigma_1^{*2} = 1$, $\sigma_2^{*2} = 1$, $\mu_1^* = 20$, and $\mu_2^* = 1$.

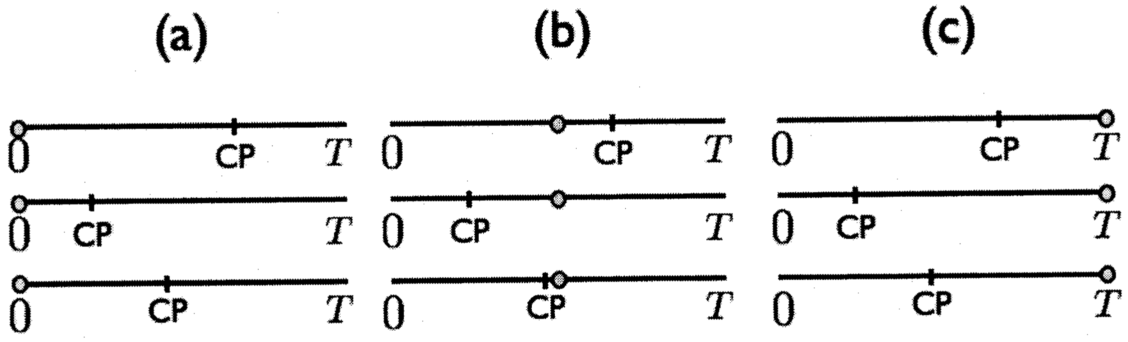


Figure 8.3: Three designs with one design point. (a) The design point is at 0. (b) The design point is in the middle of the interval $[0, T]$. (c) The design point is at T.

times, it provides the opportunity to select measurements from both the "μ_1 and μ_2 distributions".

In general, we speculate that the optimal design will occasionally place a point in the interior of $[0, T]$ to even up the number of observations from the "μ_1 and μ_2 distributions". The strategy of placing an observation in the interior of $[0, T]$ obviously provides a greater benefit when there is a large number of subjects. This claim derives from our intuition and the fact that the curvature of the convex $R_1 + R_3$ term increases with m. Note also that the common changepoint multi-path problem would not benefit from such a design because all subjects change at the same time. Hence, the number of observations coming from each of "the μ_1 and μ_2 distributions" is always a multiple of m in the common changepoint problem, regardless of the design used.

Chapter 9

Summary and Future Work

In this last chapter we provide concluding remarks concerning the thesis and consider avenues for future research.

9.1 Thesis Summary

The main results of this thesis concern Bayesian optimal designs for the single-path changepoint problem. To summarize, we found, using decision-theoretic criterion functions, that the optimal designs for testing for a change and/or estimating the before-and-after-change means is one of the designs in V placing observations as far as possible at the ends of the interval $[0, T]$, whenever the prior on the changepoint is log-concave. Likewise, the optimal design for testing for a change in the subinterval $[t_1, t_2]$ is one of the designs placing observations as close as possible towards t_1 and t_2 and possibly 0 and/or T. Since, in our model, we do not assume that the change occurs at a location where a measurement is taken, it is not possible to obtain the posterior distribution for the changepoint. Hence, we are unable to provide optimal designs for estimating the changepoint location. We stress that our optimal design results for the single-path problem apply for any distribution for the data and that conjugate priors are not needed for the before-and-after-change means.

As we saw in Section 1.2.2, design measures and convex optimization have played a large role in the subject of optimal design. Our single-path problem is no different since the design measure we introduced provided us with concave criterion functions. Using techniques from differential geometry, we were able to prove, given a log-concave prior distribution for the changepoint that we minimize a concave criterion function over a subset of a simplex, that contains the vertices of the simplex.

We concluded by considering the multiple changepoints multi-path problem. We found that even in terms of the design measure, the design criterion functions for estimating the before-and-after-change hierarchical means, are not always concave. At this point we are uncertain as to whether the design criterion functions for estimating the proportion of people who did not change and for estimating the proportion of people who changed in the subinterval $[t_1, t_2]$ are concave. We presented specific examples of the Bayes risk for estimating the hierarchical before-and-after-change means which are not concave. Hence the optimal design for the multiple changepoints multi-path problem are more complicated. Since we showed the common changepoint multi-path problem is really a single-path problem, the extra complication is due to multiple changepoints.

9.2 Future Work

We outline how our work could be extended, and suggest some other optimal design changepoint problems.

9.2.1 Extensions to the Single-Path Optimal Design Problem

Recall, that in this thesis we assume the observations to be conditionally independent. Since repeated measurements are taken on the same subject, we avoided between observation dependence by assuming that observations are taken sufficiently far apart to ensure that they are roughly conditionally independent. A major advance would be to allow for dependence between the measurements based upon the distance between them. Unfortunately, the problem then becomes much more complicated. With the added dependence of the likelihood on the design, it is unlikely the design criterion function would be concave. Intuitively, there is a "tug of war". As we have seen, when estimating the before-and-after-change means, it is optimal to take the observations far from the changepoint and, hence, it is good to crowd the observations as close as possible towards the ends of the interval. However, when there is a dependence which is a function of the distance between the measurements, we would benefit from spreading the measurements out to obtain less inter-observation dependence.

As mentioned in Section 1.1, often an instantaneous changepoint model is a good approximation to settings that display a gradual change. A natural extension would be to consider the optimal designs for problems with a gradual change. This could be done quite simply by introducing two changepoints with a change slope between them. We suspect that a multivariate design measure for the two changepoints similar to the one we proposed could be used in this problem.

In the above problems the only solutions might be through numerical optimization, with hopefully, methods for confining the search for a optimal design to a manageable sub-region of the design space.

Finally, we have observed when estimating the before-and-after-change means if the difference in the hyperparametric means is large compared to the hyperparametric and model variances, the Bayes risk can be approximated by $R_1 + R_3$. Quantifying this observation appears to be difficult because $R_2 + R_4$ is rather complicated.

9.2.2 The Multiple Changepoints Multi-Path Problem

The design criterion functions for the multiple changepoints multi-path problem do not seem to be concave in π. In fact, for estimating the before-and-after-change means, we have seen examples in which the design criterion function is definitely not concave. Therefore, we need to develop numerical techniques. One may ask whether the design measure presented in this thesis is useful for this. On the one hand, this design measure simplifies the Bayes risk. On the other hand, with this design measure, we minimize over the much more complicated region $G_f(\mathbb{X}^n)$.

9.2.3 The Biphasic Regression Problem

Many changepoint problems arising in medicine have a change in regression slope. Such a problem was introduced in the Bayesian setting by Smith and Cook (1980) who investigated renal transplants. Other examples of a change-in-slope changepoint problem are discussed by Krolewski et al. (1995), Chu et al. (2005) and Belisle et al. (2002) who investigated diabetes, AIDS, and Alzheimers' disease, respectively. Since such regression problems arise frequently and since, in a medical setting, investigators often have control over where the measurements are taken, optimal designs for such biphasic regression problems are of great value. Upon reflection, we again see that there is a "tug of war". For regular regression without a changepoint, the optimal design for estimating the slope places observations towards the ends of the regression interval. Therefore when estimating the before-and-after-change slopes we would want to place observations near the begining of the interval, near the changepoint, and near the end of the interval. However, from the results in this thesis, we know that we usually want to avoid taking measurements near the unknown changepoint location.

9.2.4 The Cross-Sectional Data Changepoint Problem

In some experiments such as IQ tests and sacrifice experiments, one cannot take repeated measurements on the same subject. Hence, each observation is taken on a different subject. Therefore, there is no need to keep measurements a distance d apart. Such problems are considered in Chapter 2 of Zhou (1997) where she modelled the observed data as coming from a mixture of two densities. Zhou (1997) proved that the optimal design will take measurements at 0 and T. This problem can be reformulated in terms of a design measure π as in the multiple changepoints multi-path problem. The Bayes risk for estimating the before-and-after-change means can again be divided into the terms R_1, R_2, R_3 and R_4. Since we do not need to keep points a minimum distance apart, we minimize over the simplex S^n. As long as we have conjugate prior distributions, we can calculate $R_1 + R_3$ analytically. We found $R_1 + R_3$ is a linear function of π and, consequently, is minimized at one of the vertices of S^n. If the term $R_2 + R_4$ is concave and zero at the vertices of S^n, the optimal design could then be found exactly from the analytical expression of $R_1 + R_3$.

Appendix A

Remarks for Chapter 5

A.1 Remarks for Theorem 5.6

Consider a single-path problem with either a Type 1, 2, or 3 log-concave prior distribution for the changepoint. We specify conditions for which we only need to consider a subset of V when using a concave design criterion function to find the optimal design.

Type 1: If $(n-k)d > (T-t) > (n-k-1)d$, we only need consider the designs $\{u_0, \ldots, u_{n-k}\}$. This is because $u_{n-k+1}, \ldots, u_n$ all map to $G_f(u_{n-k})$.

Type 2: If $(n-k)d < t < (n-k-1)d$, we consider the subset $\{u_k, \ldots, u_n\}$. This is because $u_0, \ldots, u_{k-1}$ all map to $G_f(u_k)$.

Type 3: If $(n-k-1)d < (T-t_0) < (n-k)d$ and $(n-l-1)d < t_T < (n-l)d$ then the optimal design is one of $\{u_{n-k}, \ldots, u_l\}$.

Furthermore we may impose two conditions on the Type 3 prior distribution such that log-concavity is not required.

1) If $t_T > (n-1)d$, $(T - t_0) > (n-1)d$ and $(t_T - t_0) < (n-1)d$, the optimal design is contained in V regardless of whether or not f is log-concave.

2) If $(n-k-1)d < (T - t_0) < (n-k)d$, $(n-l-1)d < t_T < (n-l)d$ and $(t_T - t_0) < (n-k-l)d$, the optimal design is one of $\{u_{n-k}, \ldots, u_l\}$ regardless of whether or not the prior distribution is log-concave.

These remarks follow almost immediately from the derivations in Sections 5.3.1, 5.3.2, and 5.3.3 .

Appendix B

Algebraic Details for Chapter 7

We present details of the common changepoint multi-path problem of Section 7.2.1.

B.1 The Density $f(y|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k)$

$$
\begin{aligned}
f(y|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k) &= \int\int f(y|m_1, m_2, \tau_x = k) f(m_1|\bar{\mu}_1) f(m_2|\bar{\mu}_2) dm_1 dm_2 \\
&= \prod_{i=1}^{m} \left(\int \prod_{j \le k} (f(y_{ij}|\mu_{1i}, \tau_x = k) f(\mu_{1i}|\bar{\mu}_1) d\mu_{1i}) \right. \\
&\times \left. \int \prod_{j > k} (f(y_{ij}|\mu_{2i}, \tau_x = k) f(\mu_{2i}|\bar{\mu}_2) d\mu_{2i}) \right)
\end{aligned} \tag{B.1}
$$

We begin by integrating $\int \prod_{j \le k} (f(y_{ij}|\mu_{1i}, \tau_x = k) f(\mu_{1i}|\bar{\mu}_1) d\mu_{1i})$ for a single subject i. The result is the same for all subjects. With $\tau_x = k$, let $y_i^1 = (y_{i1}, \dots, y_{ik})'$, the column vector of k measurements taken before the change on subject i. Recall e is a column vector of ones and I is the identity matrix.

$$
\begin{aligned}
&\prod_{j\leq k}\left(\int f(y_{ij}|\mu_{1i},\tau_x)f(\mu_{1i}|\bar{\mu}_1)d\mu_{1i}\right)\\
&=\int\left(\frac{1}{2\pi\sigma_1^2}\right)^{\frac{k}{2}}\exp\left(-\frac{1}{2}(y_i^1-\mu_{1i}e)'(\sigma_1^2I)^{-1}(y_i^1-\mu_{1i}e)\right)\\
&\times\left(\frac{1}{2\pi\bar{\sigma}_1^2}\right)^{\frac{1}{2}}\exp\left(-\frac{1}{2\bar{\sigma}_1^2}(\mu_{1i}-\bar{\mu}_1)^2\right)d\mu_{1i}\\
&=\left(\frac{1}{2\pi\sigma_1^2}\right)^{\frac{k}{2}}\left(\frac{1}{2\pi\bar{\sigma}_1^2}\right)^{\frac{1}{2}}\\
&\times\int\exp\left(-\frac{1}{2}\left(\mu_{1i}^2\left(\frac{k}{\sigma_1^2}+\frac{1}{\bar{\sigma}_1^2}\right)-2\mu_{1i}\left(\frac{\sum_{j\leq k}y_{ij}}{\sigma_1^2}+\frac{\bar{\mu}_1}{\bar{\sigma}_1^2}\right)+\frac{\bar{\mu}_1^2}{\bar{\sigma}_1^2}+\frac{\sum_{j\leq k}y_{ij}^2}{\sigma_1^2}\right)\right)d\mu_{1i}\\
&=\left(\frac{1}{2\pi\sigma_1^2}\right)^{\frac{k}{2}}\left(\frac{1}{2\pi\bar{\sigma}_1^2}\right)^{\frac{1}{2}}\left(\frac{2\pi\bar{\sigma}_1^2\sigma_1^2}{k\bar{\sigma}_1^2+\sigma_1^2}\right)^{\frac{1}{2}}\\
&\times\exp\left(-\frac{1}{2}\sum_{j\leq k}y_{ij}^2\left(\frac{(k-1)\bar{\sigma}_1^2+\sigma_1^2}{\sigma_1^2(k\bar{\sigma}_1^2+\sigma_1^2)}\right)-2\sum_{j=1}^{k}\sum_{l=1}^{j-1}y_{ij}y_{il}\left(\frac{\bar{\sigma}_1^2/\sigma_1^2}{k\bar{\sigma}_1^2+\sigma_1^2}\right)\right.\\
&\left.-2\sum_{j\leq k}y_{ij}\left(\frac{\bar{\mu}_1}{k\bar{\sigma}_1^2+\sigma_1^2}\right)+\frac{\bar{\mu}_1^2k}{k\bar{\sigma}_1^2+\sigma_1^2}\right)
\end{aligned}
\tag{B.2}
$$

Once the random effect μ_{1i} is integrated out, the covariance matrix for the the vector, y_i^1, of observations, has equal diagonal entries and equal all off-diagonal entries. Consequently, its inverse also has equal diagonal entries and equal off-diagonal entries. From the evaluation of the integral above, we identify the diagonal entries of the inverse of the covariance matrix as $\left(\frac{(k-1)\bar{\sigma}_1^2+\sigma_1^2}{\sigma_1^2(k\bar{\sigma}_1^2+\sigma_1^2)}\right)$ and the off-diagonal entries of the inverse of the covariance matrix as $-\left(\frac{\bar{\sigma}_1^2/\sigma_1^2}{k\bar{\sigma}_1^2+\sigma_1^2}\right)$.

Using the inverse covariance matrix, we can find the covariance matrix itself. The covariance matrix has size $k\times k$ and has diagonal entries all equal to $\bar{\sigma}_1^2+\sigma_1^2$ and off-diagonal entries all equal to $\bar{\sigma}_1^2$. The form of the multivariate normal density arising from the integral is $N_k(\bar{\mu}_1e,\Sigma_1)$, where Σ_1 is the covariance matrix described above. Similarly, we find that the integral, $\int\prod_{j>k}\left(f(y_{ij}|\mu_{2i},\tau_x=k)f(\mu_{2i}|\bar{\mu}_2)d\mu_{2i}\right)$ integrates to a $N_{(n-k)}(\bar{\mu}_2 e, \Sigma_2)$ density, where Σ_2 has size $(n-k) \times (n-k)$ and diagonal entries $\bar{\sigma}_2^2 + \sigma_2^2$ with off-diagonal entries $\bar{\sigma}_2^2$.

Combining these results,

$$f(y|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k) = \prod_{i=1}^{m} N_k(\bar{\mu}_1 e, \Sigma_1) N_{(n-k)}(\bar{\mu}_2 e, \Sigma_2) \tag{B.3}$$

Equation (B.3) can also be expressed in terms of the Kronecker product as shown in expression (7.31) of Section 7.2.1.

B.2 The Density $f(\bar{y}|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k)$

To find the density $f(\bar{y}|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k)$, we use the following well-know result:

Lemma B.1. *If $y \sim N_q(\mu, \Sigma)$ then $Ay \sim N_p(A\mu, A\Sigma A')$ where A is a $p \times q$ matrix.*

Here, we need a $m \times mn$ matrix A such that $Ay = \bar{y}$. From (7.21) we see, upon inspection, that A is block diagonal, with two blocks on the diagonal:

$$A = \frac{1}{m}\begin{bmatrix} B & 0 \\ 0 & C \end{bmatrix}$$

With $\tau_k = k$, the upper left block B is of size $k \times mk$ and the lower block C is of size $(n-k) \times (n-k)$. The matrix A is designed so that the rth row of A sums the rth observation of every patient. Specifically the sth row of B will have ones in the columns $lk + s$ for $l = 0, \ldots, m-1$ and a zero in every other column. Likewise the sth row of C will have ones in the columns $l(n-k)+s$ for $l = 0, \ldots, m-1$ and zeroes in every other column.

Next we use equation (7.31) to identify the μ and Σ in $y \sim N(\mu, \Sigma)$, as in Lemma B.1, and calculate $A\mu$ and $A\Sigma A'$. From equation (7.31) with $\tau_x = k$, we see that the mean vector has mk entries of $\bar{\mu}_1$ followed by $m(n-k)$ entries of $\bar{\mu}_2$. Therefore $A\mu$ has k entries of $\bar{\mu}_1$ followed by $(n-k)$ entries of $\bar{\mu}_1$. Also we can see from equation (7.31) that Σ is block diagonal with the first m blocks of size $k \times k$ having diagonal elements $\bar{\sigma}_1^2 + \sigma_1^2$ and off-diagonal elements $\bar{\sigma}_1^2$. The next m blocks have size $(n-k) \times (n-k)$, diagonal elements equal to $\bar{\sigma}_2^2 + \sigma_2^2$, and off-diagonal elements $\bar{\sigma}_2^2$. The operation reduces the $mn \times mn$ size of Σ to the $n \times n$ size of $A\Sigma A'$. This latter matrix sports two blocks. The first has size $k \times k$ and has $\frac{\bar{\sigma}_1^2+\sigma_1^2}{m}$ on the diagonals and $\frac{\bar{\sigma}_1^2}{m}$ on the off-diagonals. Call this first block $\bar{\Sigma}_1$. The second block has size $(n-k) \times (n-k)$ and has $\frac{\bar{\sigma}_2^2+\sigma_2^2}{m}$ on the diagonals and $\frac{\bar{\sigma}_2^2}{m}$ on the off-diagonals. Call this matrix $\bar{\Sigma}_2$. Therefore, the density $f(\bar{y}|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k)$ is expressible as,

$$f(\bar{y}|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k) = N_k(\bar{\mu}_1 e, \bar{\Sigma}_1) N_{(n-k)}(\bar{\mu}_2 e, \bar{\Sigma}_2). \tag{B.4}$$

B.3 Posterior Means and Variances of $\bar{\mu}_1$ and $\bar{\mu}_2$

For $\tau_x = k$, let $\bar{y}^1$ and $\bar{y}^2$ be as defined in expressions (8.4) and (8.5). From equation (B.4) we see that $f(\bar{y}|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k) = f(\bar{y}^1|\bar{\mu}_1, \tau_x = k) f(\bar{y}^2|\bar{\mu}_2, \tau_x = k)$ where $f(\bar{y}^1|\bar{\mu}_1, \tau_x = k)$ and $f(\bar{y}^2|\bar{\mu}_2, \tau_x = k)$ have $N_k(\bar{\mu}_1 e, \bar{\Sigma}_1)$ and $N_{(n-k)}(\bar{\mu}_2 e, \bar{\Sigma}_2)$ densities respectively. Starting with $f(\bar{\mu}_1|y, \tau_x = k) \propto \int f(\bar{y}|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k) f(\bar{\mu}_1) f(\bar{\mu}_2) d\bar{\mu}_2$, we find $f(\bar{\mu}_1|\bar{y}^1, \tau_x = k) \propto f(\bar{y}^1|\bar{\mu}_1, \tau_x = k) f(\bar{\mu}_1)$. Recall $f(\bar{\mu}_1)$ has a $N(\mu_1^*, \sigma_1^{*2})$ density.

$$\begin{aligned}
&f(\bar{\mu}_1|\bar{y}^1, \tau_x = k) \\
&\propto \exp\left(-\frac{1}{2}(\bar{y}^1 - \bar{\mu}_1 e)'\bar{\Sigma}_1^{-1}(\bar{y}^1 - \bar{\mu}_1 e) + \frac{\bar{\mu}_1^2}{\sigma_1^{*2}} - \frac{2\bar{\mu}_1\mu_1^*}{\sigma_1^{*2}} + \frac{\mu_1^{*2}}{\sigma_1^{*2}}\right) \\
&= \exp\left(\mu_1^2\left(e'\bar{\Sigma}_1^{-1}e + \frac{1}{\sigma_1^{*2}}\right) - \mu_1\left(e'\bar{\Sigma}_1^{-1}\bar{y}^1 + \bar{y}^{1\,\prime}\bar{\Sigma}_1^{-1}e + \frac{2\mu_1}{\sigma_1^{*2}}\right) + \left(\bar{y}^{1\,\prime}\bar{\Sigma}_1^{-1}\bar{y}^1 + \frac{\mu_1^{*2}}{\sigma_1^{*2}}\right)\right) \\
&= \exp\left(-\frac{1}{2}\left(\bar{\mu}_1^2\left(\frac{mk}{k\bar{\sigma}_1^2 + \sigma_1^2} + \frac{1}{\sigma_1^{*2}}\right) - 2\bar{\mu}_1\left(\sum_{j=1}^{k}\frac{m\bar{y}_j}{\tau_x\bar{\sigma}_1^2 + \sigma_1^2} + \frac{\mu_1^*}{\sigma_2^{*2}}\right) + \left(\bar{y}^{1\,\prime}\bar{\Sigma}^{-1}\bar{y}^1 + \frac{\mu_1^{*2}}{\sigma_1^{*2}}\right)\right)\right)
\end{aligned} \tag{B.5}$$

From the above it is easily seen that

$$E(\bar{\mu}_1|\bar{y}, \tau_x = k) = \frac{\sum_{j=1}^{k}\frac{m\bar{y}_j}{k\bar{\sigma}_1^2+\sigma_1^2} + \frac{\mu_1^*}{\sigma_1^{*2}}}{\frac{mk}{k\bar{\sigma}_1^2+\sigma_1^2} + \frac{1}{\sigma_1^{*2}}} \tag{B.6}$$

and

$$Var(\bar{\mu}_1|\bar{y}, \tau_x = k) = \frac{1}{\frac{mk}{k\bar{\sigma}_1^2+\sigma_1^2} + \frac{1}{\sigma_1^{*2}}}. \tag{B.7}$$

Similar calculations show that

$$E(\bar{\mu}_2|\bar{y}, \tau_x = k) = \frac{\sum_{j=1}^{k} \frac{m\bar{y}_j}{(n-k)\bar{\sigma}_2^2+\sigma_2^2} + \frac{\mu_2^*}{\sigma_2^{*2}}}{\frac{m(n-k)}{(n-k)\bar{\sigma}_2^2+\sigma_2^2} + \frac{1}{\sigma_2^{*2}}} \tag{B.8}$$

and

$$Var(\bar{\mu}_2|\bar{y}, \tau_x = k) = \frac{1}{\frac{m(n-k)}{(n-k)\bar{\sigma}_2^2+\sigma_2^2} + \frac{1}{\sigma_2^{*2}}}. \tag{B.9}$$

B.4 The Density $f(\bar{y}|\tau_x = k)$

Taking notational liberties, we have, using (B.4), that

$$\begin{aligned} f(\bar{y}|\tau_x = k) &= \int\int f(\bar{y}|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k) f(\bar{\mu}_1) f(\bar{\mu}_2) d\bar{\mu}_1 d\bar{\mu}_2 \\ &= \int N_k(\bar{\mu}_1 e, \bar{\Sigma}_1) f(\bar{\mu}_1) d\bar{\mu}_1 \int N_{(n-k)}(\bar{\mu}_2 e, \bar{\Sigma}_2) f(\bar{\mu}_2) d\bar{\mu}_2. \end{aligned} \tag{B.10}$$

We begin by finding $\int N_k(\bar{\mu}_1 e, \bar{\Sigma}_1) f(\bar{\mu}_1) d\bar{\mu}_1$. First, we write out the kernel of $N_k(\bar{\mu}_1 e, \bar{\Sigma}_1) f(\bar{\mu}_1)$ and collect all the $\bar{\mu}_1^2$ and $\bar{\mu}_1$ terms in expression (B.11).

$$\exp\left(-\frac{1}{2}\left(\bar{\mu}_1^2\left(\frac{1}{\sigma_1^{*2}} + e'\bar{\Sigma}_1^{-1}e\right) - \bar{\mu}_1 2\left(e'\bar{\Sigma}_1^{-1}\bar{y}^1 + \frac{\mu_1^*}{\sigma_1^*}\right) + \bar{y}^{1\prime}\bar{\Sigma}_1^{-1}\bar{y}^1 + \frac{\mu_1^{*2}}{\sigma_1^*}\right)\right) \tag{B.11}$$

Next, we integrate out $\bar{\mu}_1$. We know our final distribution will be normal for $\bar{y}^1$. After integrating $\bar{\mu}_1$ out from the kernel we find that the distribution of $\bar{y}^1$ given $\tau_x = k$ is $N_k\left(\mu_1^* e, \left(\bar{\Sigma}_1^{-1} - \frac{\bar{\Sigma}_1^{-1}ee'\bar{\Sigma}_1^{-1}}{e'\bar{\Sigma}_1^{-1}e+\frac{1}{\sigma_1^*}}\right)^{-1}\right)$. After some calculation, we find that the $k \times k$ covariance matrix, $\left(\bar{\Sigma}_1^{-1} - \frac{\bar{\Sigma}_1^{-1}ee'\bar{\Sigma}_1^{-1}}{e'\bar{\Sigma}_1^{-1}e+\frac{1}{\sigma_1^*}}\right)^{-1}$, has diagonal elements $\frac{\sigma_1^2+\bar{\sigma}_1^2}{m} + \sigma_1^{*2}$, and off-diagonal elements $\frac{\bar{\sigma}_1^2}{m} + \sigma_1^{*2}$. Similarly, we can carry out the same steps to show that $f(\bar{y}^2|\tau_x = k)$ is a $N_{(n-k)}\left(\mu_2^* e, \left(\bar{\Sigma}_2^{-1} - \frac{\bar{\Sigma}_2^{-1}ee'\bar{\Sigma}_2^{-1}}{e'\bar{\Sigma}_2^{-1}e+\frac{1}{\sigma_2^*}}\right)^{-1}\right)$ density, with the covariance matrix having diagonal elements, $\frac{\sigma_2^2+\bar{\sigma}_2^2}{m} + \sigma_2^{*2}$, and off-diagonal elements, $\frac{\bar{\sigma}_2^2}{m} + \sigma_2^{*2}$.

Appendix C

Algebraic Details for Chapter 8

C.1 The Density $f(y|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k)$

We first find $f(y|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k)$. The expression, $f(y|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k)$, is then used to calculate $f(y|\tau_x = k)$.

$$\begin{aligned} f(y|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k) &= \int\int f(y|m_1, m_2, \tau_x = k) f(m_1|\bar{\mu}_1) f(m_2|\bar{\mu}_2) dm_1 dm_2 \\ &= \prod_{i=1}^{m} \left(\int \prod_{j \le k_i} (f(y_{ij}|\mu_{1i}, \tau_{xi} = k_i) f(\mu_{1i}|\mu_1) d\mu_{1i}) \right. \\ &\times \left. \int \prod_{j > k_i} (f(y_{ij}|\mu_{2i}, \tau_{xi} = k_i) f(\mu_{2i}|\mu_2) d\mu_{2i}) \right) \end{aligned} \tag{C.1}$$

We begin by integrating $\int \prod_{j \le k_i} f(y_{ij}|\mu_{1i}, \tau_{xi} = k_i) f(\mu_{1i}|\mu_1) d\mu_{1i}$. As in Appendix B.1 we denote the k_i observations taken before the change on subject i by y_i^1. That is , $y_i^1 = (y_{i1}, \ldots, y_{ik_i})$. Letting e represent a column vector of ones of length k_i and I a

$k_i \times k_i$ identity matrix we have.

$$
\begin{aligned}
&\prod_{j\le k_i}\left(\int f(y_{ij}|\mu_{1i},\tau_{xi}=k_i)f(\mu_{1i}|\mu_1)d\mu_{1i}\right)\\
&=\int\left(\frac{1}{2\pi\sigma_1^2}\right)^{\frac{k_i}{2}}\exp\left(-\frac{1}{2}(y_i^1-\mu_{1j}e)'(\sigma_1^2I)^{-1}(y_i^1-\mu_{1i}e)\right)\\
&\times\left(\frac{1}{2\pi\bar{\sigma}_1^2}\right)^{\frac{1}{2}}\exp\left(-\frac{1}{2\bar{\sigma}_1^2}(\mu_{1i}-\bar{\mu}_1)^2\right)d\mu_{1i}\\
&=\left(\frac{1}{2\pi\sigma_1^2}\right)^{\frac{k_i}{2}}\left(\frac{1}{2\pi\bar{\sigma}_1^2}\right)^{\frac{1}{2}}\\
&\times\int\exp\left(-\frac{1}{2}\left(\mu_{1i}^2\left(\frac{k_i}{\sigma_1^2}+\frac{1}{\bar{\sigma}_1^2}\right)-2\mu_{1i}\left(\frac{\sum_{j\le k_i}y_{ij}}{\sigma_1^2}+\frac{\bar{\mu}_1}{\bar{\sigma}_1^2}\right)+\frac{\bar{\mu}_1^2}{\bar{\sigma}_1^2}+\frac{\sum_{j\le k_i}y_{ij}^2}{\sigma_1^2}\right)\right)d\mu_{1i}\\
&=\left(\frac{1}{2\pi\sigma_1^2}\right)^{\frac{k_i}{2}}\left(\frac{1}{2\pi\bar{\sigma}_1^2}\right)^{\frac{1}{2}}\left(\frac{2\pi\bar{\sigma}_1^2\sigma_1^2}{k_i\bar{\sigma}_1^2+\sigma_1^2}\right)^{\frac{1}{2}}\\
&\times\exp\Bigg(-\frac{1}{2}\sum_{j\le k_i}y_{ij}^2\left(\frac{(k_i-1)\bar{\sigma}_1^2+\sigma_1^2}{\sigma_1^2(k_i\bar{\sigma}_1^2+\sigma_1^2)}\right)-2\sum_{j=1}^{k_i}\sum_{l=1}^{j-1}y_{ij}y_{il}\left(\frac{\bar{\sigma}_1^2/\sigma_1^2}{k_i\bar{\sigma}_1^2+\sigma_1^2}\right)\\
&-2\sum_{j\le k_i}y_{ij}\left(\frac{\bar{\mu}_1}{k_i\bar{\sigma}_1^2+\sigma_1^2}\right)+\frac{\bar{\mu}_1^2k_i}{k_i\bar{\sigma}_1^2+\sigma_1^2}\Bigg)
\end{aligned}
\tag{C.2}
$$

Knowing that once the random effect μ_{1j} has been integrated out the covariance matrix for the y_i^1, has equal diagonal entries and equal off-diagonal entries implies that its inverse will also have equal diagonal entries and equal off-diagonal entries. From (C.2), we identify the diagonal entries for the inverse of the covariance matrix as $\left(\frac{(k_i-1)\bar{\sigma}_1^2+\sigma_1^2}{\sigma_1^2(k_i\bar{\sigma}_1^2+\sigma_1^2)}\right)$ and the off-diagonal entries for the inverse of the covariance matrix as $-\left(\frac{\bar{\sigma}_1^2/\sigma_1^2}{k_i\bar{\sigma}_1^2+\sigma_1^2}\right)$.

From the inverse of the covariance matrix we can find the covariance matrix itself. The covariance matrix has diagonal entries all equal to $\bar{\sigma}_1^2+\sigma_1^2$ and off-diagonal entries all equal to $\bar{\sigma}_1^2$. Hence the form of the multivariate normal density arising from the integral is $N_{k_i}(\bar{\mu}_1e,\Sigma_1)$, where Σ_1 is the covariance matrix described above. Similarly, we find that $\int\prod_{j>k_i}f(y_{ij}|\mu_{2i},\tau_{xi}=k_i)f(\mu_{2i}|\mu_2)d\mu_{2i}$ reduces to a $N_{(n-k_i)}(\bar{\mu}_2e,\Sigma_2)$ density, where Σ_2 has diagonal entries $\bar{\sigma}_2^2 + \sigma_2^2$ and off-diagonal entries, $\bar{\sigma}_2^2$. In summary, for notational brevity, we write,

$$f(y|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k) = \prod_{i=1}^{m} N_{k_i}(\bar{\mu}_1 e, \Sigma_1) N_{(n-k_i)}(\bar{\mu}_2 e, \Sigma_2). \tag{C.3}$$

C.2 Posterior Means and Variances of $\bar{\mu}_1$ and $\bar{\mu}_2$

Starting with $f(\bar{\mu}_1|y^1, \tau_x = k) \propto \int f(y^1|\bar{\mu}_1, \bar{\mu}_2, \tau_x = k) f(\bar{\mu}_1) f(\bar{\mu}_2) d\bar{\mu}_2$ we find $f(\bar{\mu}_1|y^1, \tau_x = k) \propto f(y^1|\bar{\mu}_1, \tau_x = k) f(\bar{\mu}_1)$ where $f(y^1|\bar{\mu}_1, \tau_x = k) = \prod_{i=1}^{m} N_{k_i}(\bar{\mu}_1 e, \Sigma_1)$ and $f(\bar{\mu}_1) = N(\mu_1^*, \sigma_1^{*2})$. Therefore, we find

$$\begin{aligned}
&f(\bar{\mu}_1|y^1, \tau_x = k) \\
&\propto \exp\left(-\frac{1}{2}(y^{1\,\prime} - \bar{\mu}_1 e)'\Sigma_1^{-1}(y^1 - \bar{\mu}_1 e) + \frac{\bar{\mu}_1^2}{\sigma_1^{*2}} - \frac{2\bar{\mu}_1\mu_1^*}{\sigma_1^{*2}} + \frac{\mu_1^{*2}}{\sigma_1^{*2}}\right) \\
&= \exp\left(\bar{\mu}_1^2\left(e'\Sigma_1^{-1}e + \frac{1}{\sigma_1^{*2}}\right) - \mu_1\left(e'\Sigma_1^{-1}y^1 + y^{1\,\prime}\Sigma_1^{-1}e + \frac{2\mu_1}{\sigma_1^{*2}}\right) + \left(y^{1\,\prime}\Sigma_1^{-1}y^1 + \frac{\mu_1^{*2}}{\sigma_1^{*2}}\right)\right) \\
&= \exp\left(-\frac{1}{2}\left(\bar{\mu}_1^2\left(\sum_{i=1}^{m}\frac{k_i}{k_i\bar{\sigma}_1^2 + \sigma_1^2} + \frac{1}{\sigma_1^{*2}}\right) - 2\bar{\mu}_1\left(\sum_{i=1}^{m}\frac{\sum_{j\le k_i} y_{ij}}{k_i\bar{\sigma}_1^2 + \sigma_1^2} + \frac{\mu_1^*}{\sigma_2^{*2}}\right) + \left(y^{1\,\prime}\Sigma^{-1}y^1 + \frac{\mu_1^{*2}}{\sigma_1^{*2}}\right)\right)\right)
\end{aligned} \tag{C.4}$$

From (C.4), we see that,

$$E(\bar{\mu}_1|y, \tau_x = k) = \frac{\sum_{i=1}^{m}\frac{\sum_{j\le k_i} y_{ij}}{k_i\bar{\sigma}_1^2+\sigma_1^2} + \frac{\mu_1^*}{\sigma_1^{*2}}}{\sum_{i=1}^{m}\frac{k_i}{k_i\bar{\sigma}_1^2+\sigma_1^2} + \frac{1}{\sigma_1^{*2}}} \tag{C.5}$$

and

$$Var(\bar{\mu}_1|y, \tau_x = k) = \frac{1}{\sum_{i=1}^{m}\frac{k_i}{k_i\bar{\sigma}_1^2+\sigma_1^2} + \frac{1}{\sigma_1^{*2}}} \tag{C.6}$$

Similarly, for $\bar{\mu}_2$ we find,

$$E(\bar{\mu}_2|y, \tau_x = k) = \frac{\sum_{i=1}^{m}\frac{\sum_{j>k_i} y_{ij}}{(n-k_i)\bar{\sigma}_2^2+\sigma_2^2} + \frac{\mu_1^*}{\sigma_2^{*2}}}{\sum_{i=1}^{m}\frac{(n-k_i)}{(n-k_i)\bar{\sigma}_2^2+\sigma_2^2} + \frac{1}{\sigma_2^{*2}}} \tag{C.7}$$

and

$$Var(\bar{\mu}_2|y,\tau_x=k)=\frac{1}{\sum_{i=1}^{m}\frac{(n-k_i)}{(n-k_i)\bar{\sigma}_2^2+\sigma_2^2}+\frac{1}{\sigma_2^{*2}}} \tag{C.8}$$

C.3 The Density $f(y|\tau_x = k)$

$$\begin{aligned} f(y|\tau_x=k) &= \int\int f(y|\bar{\mu}_1,\bar{\mu}_2,\tau_x=k)f(\bar{\mu}_1)f(\bar{\mu}_2)d\bar{\mu}_1 d\bar{\mu}_2 \\ &= \int\prod_{i=1}^{m}N_{k_i}(\bar{\mu}_1 e,\Sigma_1)f(\bar{\mu}_1)d\bar{\mu}_1\int\prod_{i=1}^{m}N_{(n-k_i)}(\bar{\mu}_2 e,\Sigma_2)f(\bar{\mu}_2)d\bar{\mu}_2 \end{aligned} \tag{C.9}$$

We start by finding $\int\prod_{i=1}^{m}N_{k_i}(\bar{\mu}_1 e,\Sigma_1)f(\bar{\mu}_1)d\bar{\mu}_1$. First, we re-express the term $\prod_{i=1}^{m}N_{k_i}(\bar{\mu}_1 e,\Sigma_1)f(\bar{\mu}_1)$ in vector form to simply our calculations. Let M be the diagonal matrix with the first k_1 entries $\frac{1}{k_1\bar{\sigma}_1^2+\sigma_1^2}$, the next k_2 entries $\frac{1}{k_2\bar{\sigma}_1^2+\sigma_1^2}$, etc. Finally, let E be a block diagonal matrix with m blocks. The ith block is of size $k_i \times k_i$ and has entries $\frac{\bar{\sigma}_1^2/\sigma_1^2}{k_i\bar{\sigma}_1^2+\sigma_1^2}$.

$$\prod_{i=1}^{m}\frac{(\sigma^2)^{\frac{1}{2}}}{(2\pi)^{\frac{k_i}{2}}|\sigma_1^2 I|(\bar{\sigma}_1^2 k_i+\sigma_1^2)^{\frac{1}{2}}(2\pi)^{\frac{1}{2}}}$$
$$\times\exp\left(-\frac{1}{2}\left(\bar{\mu}_1^2\left(\sum_{i=1}^{m}\frac{k_i}{k_i\bar{\sigma}_1^2+\sigma_1^2}+\frac{1}{\sigma_1^{*2}}\right)-2\bar{\mu}_1\left(y^1{}'Me+\frac{\mu_1^*}{\sigma_1^{*2}}\right)+y^1{}'(\sigma_1^2 I)^{-1}y^1-y^1{}'Ey^1+\frac{\mu_1^{*2}}{\sigma_1^{2*}}\right)\right)$$

Integrating out $\bar{\mu}_1$, this expression becomes

$$\prod_{i=1}^{m}\frac{(\sigma_1^2)^{\frac{1}{2}}(2\pi)^{\frac{1}{2}}\left(\frac{1}{\sum_{i=1}^{m}\frac{k_i}{k_i\bar{\sigma}_1^2+\sigma_1^2}+\frac{1}{\sigma_1^{2*}}}\right)^{\frac{1}{2}}}{(2\pi)^{\frac{k_i}{2}}|\sigma_1^2 I|(\bar{\sigma}_1^2 k_i+\sigma_1^2)^{1/2}(2\pi\sigma_1^{*2})^{\frac{1}{2}}}$$
$$\times\exp\left(-\frac{1}{2}\left(y^1{}'\left((\sigma_1^2 I)^{-1}-E-\frac{Me'eM}{\sum_{i=1}^{m}\frac{k_i}{k_i\bar{\sigma}_1^2+\sigma_1^2}+\frac{1}{\sigma_1^{2*}}}\right)y^1+y^1{}'\frac{2M\frac{\mu_1^*}{\sigma_1^{*2}}}{\sum_{i=1}^{m}\frac{k_i}{k_i\bar{\sigma}_1^2+\sigma_1^2}+\frac{1}{\sigma_1^{2*}}}e+\frac{\mu_1^*}{\sigma_1^{*2}}\right)\right) \tag{C.10}$$

Since the inverse of the covariance matrix is $\left((\sigma_1^2 I)^{-1}-E-\frac{Me'eM}{\sum_{i=1}^{m}\frac{k_i}{k_i\bar{\sigma}_1^2+\sigma_1^2}+\frac{1}{\sigma_1^{2*}}}\right)$, the covariance matrix has a form with equal diagonal entries, $\sigma_1^2+\bar{\sigma}_1^2+\sigma_1^{*2}$. There is a block diagonal component to the matrix with blocks of size k_i for $i = 1, \ldots, m$. The off-diagonal components in these blocks are $\bar{\sigma}_1^2 + \sigma_1^{*2}$. All other entries in the covariance matrix are σ_1^{*2}. Therefore with e as a column vector of ones and Σ_1^* the covariance matrix described above, we have found that $y^1 \sim N_{(\sum_{i=1}^m k_i)}(\mu_1^* e, \Sigma_1^*)$. Similarly, with Σ_2^* the same form as Σ_1^* but with blocks of size $n - k_i$, and entries with subscripts 2 instead of 1 we find that $y^2 \sim N_{(\sum_{i=1}^m (n-k_i))}(\mu_2^* e, \Sigma_2^*)$. The product of these densities describes the posterior density $f(y|\tau_x = k)$.

$$f(y|\tau_x = k) = N_{(\sum_{i=1}^m k_i)}(\mu_1^* e, \Sigma_1^*) N_{(\sum_{i=1}^m (n-k_i))}(\mu_2^* e, \Sigma_2^*) \tag{C.11}$$

Note that, even if taken on different subjects observations are correlated if either they are both taken before the change or both taken after the change. This correlation is induced by the common hierarchical means $\bar{\mu}_1$ and $\bar{\mu}_2$, respectively.

Bibliography

Atkinson, A. and Donev, A. (1992). *Optimum Experimental Designs.* Clarendon Press, Oxford UK.

Atkinson, A. and Fedorov, V. (1975). Optimal design: Experiments for discriminating between several models. *Biometrika*, 62(2):289–303.

Bagnoli, M. and Bergström, T. (1989). Log-concave probability and its applications. Manuscript, University of Michigan.

Barndorff-Nielsen, O. (1978). *Information and exponential families in statistical theory.* John Wiley & Sons.

Beckage, B., Joseph, L., Belisle, C., Wolfson, D., and Platt, W. (2006). Bayesian changepoint analyses in ecology. *The New Phytologist.*

Belisle, P., Joseph, L., Wolfson, D., and Zhou, X. (2002). Bayesian estimation of cognitive decline in patients with Alzheimer's disease. *Canadian Journal of Statistics*, 30(1):37–54.

Ben-Tal, A. and Nemirovskii, A. (2001). *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications (MPS-SIAM Series on Optimization).* Society for Industrial & Applied Math.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis.* Spinger Series in Statistics.

Bernardo, J. and Smith, A. (1994). *Bayesian Theory.* Wiley.

Bhattacharya, G. and Johnson, R. (1968). Non-parametric tests for shift at an unknown time point. *Annals of Mathematical Statistics*, 39:1731–1743.

Bischoff, W. and Miller, F. (2000). Asymptotically optimal tests and optimal designs for testing the mean in regression models with applications to change-point problems. *Annals of the Institute of Statistical Mathematics*, 52:658–679.

Blackmore, L. and Williams, B. (2005). Finite horizon control design for optimal model discrimination. In *Proceedings of the Joint 44th IEEE Conference on Decision and Control and European Control Conference.*

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3.

Brooks, R. (1972). A decision theory approach to optimal regression designs. *Biometrika*, 59:563–571.

Brooks, R. (1974). On the choice of an experiment for prediction in linear regression. *Biometrika*, 61(2):303–311.

Brooks, R. (1976). Optimal regression designs for prediction when prior knowledge is available. *Metrika*, 23:221–230.

Carlin, B., Gelfand, A., and Smith, A. (1992). Hierarchical Bayesian analysis of changepoint problems. *Applied Statistics*, 41(2):389–405.

Carter, R. and Blight, B. (1981). A Bayesian change-point problem with an application to the prediction and detection of ovulation in women. *Biometrics*, 37(4):743–751.

Casella, G. and Berger, R. (2002). *Statistical Inference Second edition.* Duxbury.

Chaloner, K. (1984). Optimal Bayesian experimental design for linear models. *Annals of Statistics*, 12 (1):283–300.

Chaloner, K. and Larntz, K. (1989). Optimal Bayesian design applied to logistic regression experiments. *Journal of Statistical Planning and Inference*, 21:191–208.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, 10:273–310.

Chen, J. and Gupta, A. (2000). *Parametric Statistical Change Point Analysis.* Bikhauser.

Chernoff, H. and Zacks, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *Annals of Mathematical Statistics*, 35:999–1018.

Christensen, J. and Rudemo, M. (1998). Multiple change-point analysis applied to the monitoring of salmonella prevalence in danish pigs and porc. *Preventive Veterinary Medicine*, 36(2):131–143.

Chu, H., Gange, S., Yamashita, T., Hoover, D., Chmiel, J., Margolick, J., and Jacobson, L. (2005). Individual variation in cd4 cell count trajectory among human immunodeficiency virus-infected men and women on long-term highly active antiretroviral therapy: An application using a Bayesian random change-point model. *American Journal of Epidemiology*, 162(8):787–797.

Chu, P. and Zhao, X. (2004). Bayesian change-point analysis of tropical cyclong activity: The central north pacific case. *American Meterological Society*, 17:4893–4901.

Clyde, M. (2001). Experimental design: A Bayesian perspective. Technical report, Duke University.

Clyde, M. and Chaloner, K. (1996). The equivalence of constrained and weighted designs in multiple objective design problems. *Journal of the American Statisitcal Association*, 91(435):1236–1244.

Consonni, G. and Veronese, P. (1992). Conjugate priors for exponential families having quadratic variance functions. *Journal of the American Statistical Association, Theory and Methods*, 87(420):1123–1127.

Cook, R. and Wong, W. (1994). On the equivalence of constrained and compound optimal design. *Journal of the American Statistical Association*, 89:687–692.

Csörgö, M. and Horváth, L. (1998). *Limit Theorems in Change-Point Analysis.* Wiley.

DeFinetti, B. (1962). Does it make sense to speak of good probability appraisers? In Good, I., Mayne, A., and Smith, J., editors, *The Scientist Speculates (An Anthology of Partly-Baked Ideas)*, pages 357–364. Capricorn Books.

Dette, H. (1993). Elfving's theorem for d-optimality. *Annals of Statistics*, 21:753–766.

Dette, H. and O'Brien, T. (1999). Optimality criteria for regression models based on predicted varience. *Biometrika*, 86(1):93–106.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *Annals of Statistics*, 7 (2):261–281.

El-Krunz, S. and Studden, W. (1991). Bayesian optimal design for linear regression models. *Annals of Statistics*, 19(4):2183–2208.

Elfving, G. (1952). Optimum allocation in linear regression theory. *Annals of Mathematical Satistics*, 23(2):255–262.

Fedorov, V. (1972). *Theory of Optimal Experiments.* Academic Press; New York.

Felsenstein, K. (1992). Optimal Bayesian design for discrimination among rival models. *Computational Statistics and Data Analysis*, 14(4):427–436.

Ghorbanzadeh, D. and Lounes, R. (2001). Bayesian analysis for detecting a change in exponential family. *Applied mathematics and computation*, 124(1):1–15.

Gneiting, T. and Raftery, A. (2004). Strictly proper scoring rules, prediction, and estimation. Technical report, University of Washington.

Goos, P., Tack, L., and Vandebroek, M. (2005). Optimal design of blocked experiments in industry. In Berger, M. and Wong, W., editors, *Applied Optimal Designs*, pages 247–279. Wiley.

Gutiérrez-Pena, E. and Smith, A. (1995). Conjugate parameterizations for natural exponential families. *Journal of the American Statistical Association*, 90(432):1347–1356.

Gutiérrez-Pena, E. and Smith, A. (1997). Exponential and Bayesian conjugate families: Review and extensions. *Test*, 6(1):1–90.

Hawkins, D. M. (1977). Testing a sequence of observations for a shift in location. *Journal of the American Statistical Association*, 72(357):180–186.

Henderson, R. and Matthews, J. (1993). An investigation of changepoints in the annual number of cases of haemolytic uraemic syndrome. *Journal of the Royal Statistical Society Series A*, 42(3):461–471.

Hinkely, D. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, 57(1):1–17.

Hinkley, D. and Hinkley, E. (1970). Inference about the change-point in a sequence of binomial variables. *Biometrika*, 57(3):477–488.

Jorgensen, B. (1997). *The Theory of Dispersion Models.* Chapman & Hall/CRC.

Joseph, L. (1990). *The Multi-Path Change-Point.* PhD thesis, McGill University.

Joseph, L. and Wolfson, D. (1992). Estimation in multi-path changepoint problems. *Comm. Statist. Theory and Methods*, 21(4):897–913.

Joseph, L. and Wolfson, D. (1993). Maximum likelihood estimation in the multipath change-point problem. *Annals of the Institute of Statistical Mathematics*, 45(3):511–530.

Joseph, L., Wolfson, D., Blisle, P., Brooks, J., Mortimer, J., Tinklenberg, J., and Yesavage, J. (1999). Taking account of between-patient variability when modeling decline in alzheimer's disease. *American Journal of Epidemiology*, 149(10):963–973.

Joseph, L., Wolfson, D., du Berger, R., and Lyle, R. (1996). Change-point analysis of a randomized trial on the effects of calcium supplementation on blood pressure. In Berry, D. and Stangl, D., editors, *Bayesian Biostatistics*, pages 617–649. Marcel Dekker, New York.

Joseph, L., Wolfson, D., du Berger, R., and Lyle, R. (1997). Analysis of panel data with change-points. *Statistical Sinica*, 7(3):687–703.

Kiefer, J. (1958). On the nonrandomized optimality and randomized nonoptimality of symmetrical designs. *Annals of Mathematical Statistics*, 29(3):675–699.

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *Annals of Statistics*, 2(5):849–879.

Kiefer, J. and Wolfowitz, J. (1959). Optimum designs in regression problems. *Annals of Mathematical Statistics*, 30(2):271–294.

Kiefer, J. and Wolfowitz, J. (1960). The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366.

Krolewski, A., Laffel, L., Krolewski, M., Quinn, M., and Warram, J. (1995). Glycosylated hemoglobin and the risk of microalbuminuria in patients with insulin-dependent diabetes mellitus. *New England Journal of Medicine*, 332(19):1251–1255.

Lauter, E. (1974). Experimental designs in a class of models. *Math. Operat. Statist.*, 5(4/5):379–396.

Letac, G. (1992). *Lectures on Natural Exponential Families and their Variance Functions.* Instituto de matematica pura e aplicada.

Lindley, D. (1956). On the measure of information provided by an experiment. *Annals of Statistics*, 27:986–1005.

Lindley, D. (1968). The choice of variables in multiple regression. *Journal of Royal Statistical Society. Series B (Methodological)*, 30(1):31–66.

Lindley, D. (1972). *Bayesian Statistics, A Review.* SIAM Philadelphia, PA.

Lyle, R., Melby, C., Hyner, G., Edmonson, J., Miller, J., and Weinberger, M. (1987). Blood pressure and metabolic effects of calcium supplementation in normotensive white and black men. *Journal of the American Medical Association*, 257(13):1772–1776.

Matthews, J. (1987). Optimal crossover designs for the comparison of two treatments in the presence of carryover effects and autocorrelated errors. *Biometrika*, 74(2):311–320.

Mira, A. and Petrone, S. (1994). Bayesian hierarchical nonparametric inference for changepoint problems. In *Bayesian Statistics 5: Proceedings of the Fifth Valencia International Meeting.*

Morris, C. (1982). Natural exponential families with quadratic variance functions. *Annals of Statistics*, 10 (1):65–80.

Morris, C. (1983). Natural exponential families with quadratic variance functions: Statistical theory. *Annals of Statistics*, 11 (2):515–529.

Muliere, P. and Scarsini, M. (1985). Change-point problems: a Bayesian nonparametric approach. *Appl. Mat.*, 30:397–402.

Müller, H. and Wang, J. (1994). Change-point problems. In Carlsten, E. E., Mller, H., and Siegmund, D., editors, *Change-point models for hazard functions*, pages 224–241. IMS Lecture Notes and Monograph Series 23.

O'Brien, T. and Funk, G. (2003). A gentle introduction to optimal design for regression models. *The American Statistician*, 57(4):265–267.

O'Neil, B. (1966). *Elementary Differential Geometry.* Academic Press.

Page, E. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2):100–14.

Page, E. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):523–7.

Page, E. (1957). On problems in which a change in parameter occurs at an unknown point. *Biometrika*, 44(1/2):248–52.

Pazman, A. (1986). *Foundations of Optimum Experimental Design.* D. Reidel Publishing Company; Dordrecht.

Pettitt, A. (1979). A non-parametric approach to the change-point problem. *Applied Statistics*, 28(2):126–135.

Picard, D. (1985). Testing and estimating change-points in time series. *Advances in Applied Probability*, 17(4):841–867.

Pilz, J. (1991). *Bayesian Estimation and Experimental Design in Linear Regression Models.* John Wiley & Sons; Bergakademie Freiberg.

Pukelsheim, F. (1993). *Optimal Design of Experiments.* John Wiley & Sons, Inc; New York.

Robert, C. (2001). *The Bayesian Choice.* Spinger Texts in Statistics.

Rockafellar, R. (1970). *Convex Analysis.* Princeton University Press, New Jersey.

Rukhin, A. (1988). Loss functions for loss estimation. *Annals of Statistics*, 16(3):1262–1269.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:279–423, 623–656.

Shewhart, W. (1931). *Economic control of quality of manufactured product.* Macmillan; New York.

Slivery, S. (1980). *Optimal Design (An Introduction to the Theory of Parameter Estimation).* Chapman and Hall.

Smith, A. (1975). A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, 62(2):407–416.

Smith, A. and Cook, D. (1980). Straight lines with a change-point: a Bayesian analysis of some renal transplant data. *Applied Statistics*, 29(2):180–189.

Spezzaferri, F. (1988). Nonsequential designs for model discrimination and parameter estimation. In Bernardo, J., DeGroot, M., and Lindley, D., editors, *Bayesian Statistics 3*, pages 777–783. Oxford University Press.

Thorpe, J. (1979). *Elementary Topics in Differential Geometry (Undergraduate Texts in Mathematics).* Springer-Verlag.

Whittle, P. (1973). Some general points in the theory of optimal experimental design. *Journal of the Royal Statistical Society. Series B (Methodological)*, 35(1):123–130.

Worsley, K. (1986). Confidence regions and tests for a changepoint in a sequence of exponential family random variables. *Biometrika*, 73(1):91–104.

Wu, Y. (2005). *Inference for Changepoint and Post Change Means After a CUSUM Test (Lecture Notes in Statistics).* Spinger.

Xu, H. (1999). Universally optimal design for computer experiments. *Statistica Sinica*, 9:1083–1088.

Zacks, S. (1983). Survey of classical and Bayesian approaches to the change-point problem: Fixed sample and sequential procedures of testing and estimation. In *Recent Advances in Statistics (Papers in Honor of Herman Chernoff on his Sixtieth Birthday).* Academic Press.

Zhou, X. (1997). *Optimal Design for Change Point Problems.* PhD thesis, McGill University.

Zhou, X., Joseph, L., Wolfson, D., and Belisle, P. (2003). A Bayesian A-optimal and model robust design criterion. *Biometrics*, 59(4):1082–1088.