UNIVERSITY OF SOUTHAMPTON
FACULTY OF SOCIAL AND HUMAN SCIENCES
School of Mathematics

Design of Experiments with Mixed Effects and Discrete Responses plus Related Topics

Timothy William Waite
M.A., M.Sc.

Thesis for the degree of Doctor of Philosophy

November 2012

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL AND HUMAN SCIENCES

School of Mathematics

Doctor of Philosophy

DESIGN OF EXPERIMENTS WITH MIXED EFFECTS AND DISCRETE RESPONSES, PLUS RELATED TOPICS

By Timothy William Waite

For certain types of experiment, the response cannot be adequately modelled using a normal distribution. When this is the case, it is common to use a Generalised Linear Model (GLM) to analyse the data. Such models allow us to fit a wide range of response distributions including Bernoulli and Poisson.

If responses in the same block are correlated, it may be appropriate to model the impact of blocking using random effects. The GLM can be extended in several ways to include random effects; both Generalised Linear Mixed Models (GLMMs) and Hierarchical Generalised Linear Models (HGLMs) are common examples of such extensions. Another example is a random intercept model for a binary response bioassay study with repeated measurements on heterogeneous individuals. The latter model is related to a GLMM but not strictly within that class.

Obtaining designs for non-normal models with random effects is complicated by the fact that the information matrix, on which most optimality criteria are based, is computationally expensive to evaluate. Indeed, if one computes naively, the search for a typical optimal GLMM design is likely to take several months.

When estimating GLMMs, it is common to use analytical approximations such as marginal quasi-likelihood (MQL) and penalised quasi-likelihood (PQL) in place of full maximum likelihood estimation. In Chapters 2 and 3, we consider the use of such computationally cheap approximations to construct surrogates for the information matrix when producing optimal designs. These reduce the computational burden substantially, enabling us to obtain designs within a practical time frame. The accuracy of the analytical approximations is explored through the use of a detailed computational approximation, which enables us to compute the optimal maximum likelihood design in the case where there are at most two points per block. It is found that one of the analytical approximations appears to perform consistently better than the others for the purposes of producing designs.

In Chapters 4 and 5, designs for an individual variation bioassay model are obtained in the cases where (i) there is a single observation, or (ii) there are multiple observations, per individual. In the former case, designs on the basis of both maximum likelihood and analytical approximations are found and compared. In the multiple observation case, a restriction on the design space enables optimal designs to be computed using a computational approximation related to that for GLMMs. This involves extensive precomputation of numerical integrals.

In Chapter 6 designs for HGLMs are studied using a computationally inexpensive asymptotic approximation to the variance-covariance matrix of the parameter estimators. This allows us to derive designs which are also efficient for the estimation of the random effects.

Throughout, the dependence of the optimal design on the unknown values of the model parameters is addressed through the use of Bayesian methods, which codify uncertainty about the parameter values using a prior distribution. We often assess the performance of the designs obtained from the optimisation of a Bayesian objective function in terms of the distribution on the local efficiencies which is induced by the prior distribution.

When the parameter space contains degenerate values, there is a problem with potential non-convergence of the Bayesian objective function used to select designs. This issue is explored in depth in Chapter 7, and results are obtained for a number of standard models.

Contents

II Dose-response experiments with unit variation 87

List of Tables

List of Figures

Declaration of authorship

I, Timothy William Waite, declare that the thesis entitled *'Design of Experiments with Mixed Effects and Discrete Responses, plus Related Topics'*, and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has been previously submitted for a degree or any other qualification or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all the main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of Chapters 2 & 3 have been published as a technical report in the Southampton Statistical Sciences Research Institute preprint series, as Waite, Woods and Waterhouse (2012).

Signed:

Date: 5 November, 2012

Acknowledgements

I would like to thank my supervisor, Dr. Dave Woods of Southampton Statistical Sciences Research Institute (S^{3}RI), for his guidance and enthusiasm at all stages of this work.

In addition, I am grateful to Dr. Tim Waterhouse[a], of Eli Lilley, USA, and Dr. Peter Van de Ven[b], of VU University Medical Center, Amsterdam, for informative discussions on GLMMs[a] and HGLMs[b] prior to embarking on the work resulting in Chapters 2 and 6 respectively.

This work was supported by studentship funding from the UK Engineering and Physical Sciences Research Council, S^{3}RI and the School of Mathematics, University of Southampton.

For their emotional support throughout, I am indebted to my parents, Bill and Trish, and my girlfriend, Hannah.

Chapter 1

Introduction

1.1 Background

In science, experiments are regularly conducted to investigate the impact of changing the values of controllable explanatory variables, labelled $x_1, \ldots, x_q$, on a response y. Such experiments consist of applying *treatments*, that is combinations of particular values of the x_i, to experimental units and observing the resulting values of y. By choosing a good experimental design, in other words making a wise selection of the treatments to be applied, the quality of the inference about the effect of the x_i can be greatly improved. The subject of design of experiments within Statistics is concerned with the common case where, despite controlling the factors $x_1, \ldots, x_q$, there remains uncontrolled variation in the response. In this scenario, the data collected are best interpreted via statistical analyses involving the fitting of stochastic models. Of prime importance therefore are the statistical properties of the inference to be drawn, particularly insofar as these depend on the choice of experimental design. In this thesis, we focus on design for classes of models which accommodate two main features, namely (i) non-normality of the response, and (ii) grouping of the experimental units into blocks. We give details of these features below. For these models, which are relatively complex, the calculation of the statistical properties of a given design involves a nontrivial amount of computation.

Non-normality of response. Much effort has been devoted to developing theory and methods for design under the assumption that the response follows a normal (Gaussian) distribution. Notable examples include the classic topics of factorial designs and their fractions, as well as response surface methodology (e.g. Myers, Montgomery and Anderson-Cook, 2009). These methods have been applied extensively in areas such as agriculture and the chemical industry. However in certain applications the response y cannot be modelled adequately using a normal distribution, making it necessary to use more sophisticated statistical tools than the linear model. For instance, in some bioassays and reliability tests the outcome measure is binary, taking values 0 or 1 only. Binary responses also feature in the aeronautical industry experiment discussed by Woods and Van de Ven (2011). Here, the outcome of interest was whether or not a spray-coating contained cracks following its application to an engine bearing. In other industrial situations the response is a continuous, positive random variable best modelled using a Gamma distribution (Robinson, Myers and Montgomery, 2004; Robinson, Wulff, Montgomery and Khuri, 2006). Alternatively, the outcome may be a count, in which case a Poisson distribution may be appropriate. A wide range of non-normal distributions for the response is available through the framework of generalised linear models (GLMs; McCullagh and Nelder, 1989). The design problem for GLMs has been discussed, among others, by Chaloner and Larntz (1989), Woods, Lewis, Eccleston and Russell (2006) and Russell, Woods, Lewis and Eccleston (2009). For more details of designs for GLMs and their extensions, see the literature review in Section 1.2.

Blocking. The experimental units in the examples of Woods and Van de Ven (2011) and Robinson et al. (2004, 2006) are grouped into homogeneous sets, called blocks, within which the responses are correlated. This feature often arises in experiments on manufacturing processes where there are typically batch effects. The units within a particular batch may be regarded as a block. Grouping of experimental units also occurs in biological or clinical experiments since repeated observations on an individual tend to be correlated. Thus the collection of measurements on a given individual constitutes a block. The statistical model and the experimental design should take into account the potential effect of these blocks. In particular, when selecting a blocked experimental design, in addition to the choice of treatments one must consider also the division of the treatments among the different blocks.

When performing a regression analysis of the data, one way to take into account the impact of blocks is to include extra terms in the predictor. Another approach is to directly model the correlation between responses in the same block. For some details on the second approach, see Section 1.2.4. We concentrate mainly on the inclusion of additional terms in the predictor. In this case, one has a choice whether to use either *fixed effects* or *random effects* for blocks. For a detailed discussion on the distinction, see McCulloch and Searle (2001, Chapter 1). Random effects are the most appropriate choice when the blocks can be regarded as a sample which is drawn from a wider population. We believe that this is usually the case in the industrial and biological examples mentioned above, and as a result we focus on models with random block effects for the majority of the thesis. The foremost advantage of using a random effects model is that it is possible to make predictions about the response for blocks other than the ones that were actually present in the experiment, i.e. future batches or future patients. Random block effects are also important in split-plot experiments, where it is not possible to estimate simultaneously fixed effects for both blocks and whole-plot factors (for further details, see Chapter 6). These advantages do however come at the expense of an additional level of parametric modelling assumptions: to implement the strategy we usually assume that the random effects are drawn from a normal distribution with mean 0. There are two main classes of models for non-normal data which also incorporate random effects. These are generalised linear mixed models (GLMMs; McCulloch and Searle, 2001) and hierarchical generalised linear models (HGLMs; Lee, Nelder and Pawitan, 2006). We discuss design for these models in Chapters 2–3 and 6 respectively.

Let $\mathbf{x} = (x_1, \ldots, x_q)^T \in \mathcal{X}$, where $\mathcal{X}$ is the *design space* of possible treatments. From a mathematical viewpoint, in this thesis a *design* is in general a measure on $\mathcal{B} = \mathcal{X}^m$, $m \geq 1$, where the term measure is used in the measure-theoretic sense (e.g. Billingsley, 2012). The space $\mathcal{B}$ corresponds to the set of combinations of m treatment vectors which can be applied together within a block of m experimental units. Mostly our design measures will have a finite support and so they will be discrete measures, although on some occasions in Chapter 7 designs will be defined by probability density functions (i.e. the measures are absolutely continuous). An *approximate* design is a probability measure, and does not correspond to a particular sample size. In the finitely-supported case, being a probability measure means that the design ξ can be written as

$$\xi = \left\{ \begin{array}{ccc} \zeta_1 & \dots & \zeta_b \\ w_1 & \dots & w_b \end{array} \right\}, \tag{1.1}$$

with $\zeta_k \in \mathcal{B}$, $w_k > 0$, $1 \leq k \leq b$, and $\sum_{k=1}^{b} w_k = 1$. The weight w_k is interpreted as the approximate proportion of available experimental blocks (assumed to be of size m) which should receive the particular combination of treatments ζ_k. Clearly to implement a general approximate design for a particular sample size, n, the weights must be rounded since it may not be the case that nw_k is an integer. For rounding procedures which result in efficient designs, see Pukelsheim and Rieder (1992). We focus on approximate designs in all of the thesis except Chapter 6, where we investigate exact designs. An *exact* design of size n is a finitely-supported counting measure, ξ', on $\mathcal{B}$. The form (1.1) still applies but now instead the w_k are positive integers which must sum to n. The interpretation here is simpler: w_k gives the precise number of blocks which should use the treatments given in ζ_k. From our perspective, the unblocked designs and analysis more commonly encountered in the optimal design literature (e.g. Atkinson, Donev and Tobias, 2007) correspond to the case $m = 1$, together with a model which does not incorporate block heterogeneity.

Optimality criteria. The optimal design paradigm is to find an approximate or exact design, ξ, which optimises the value of an objective function. This function should be chosen to reflect the purpose of experimentation. Atkinson et al. (2007) provide an extensive introductory overview, with many references to research papers providing further technical details. The advantage of this approach is that the resulting designs are tailored to specific problems, thereby increasing the efficiency of estimation and inferences drawn. The most common optimality criteria in the literature relate to the variance of estimators of the model parameters. For instance, a D-optimal design minimises the determinant of the asymptotic variance-covariance matrix of the parameter estimators (Atkinson et al., 2007, Ch. 11) and thus yields, in a sense, asymptotically optimal point estimates. We use variants of the D-criterion to select designs in this thesis. If maximum likelihood is used to estimate the parameters, $\boldsymbol{\theta} \in \mathbb{R}^p$, then the asymptotic estimator variance is proportional to $M(\xi;\boldsymbol{\theta})^{-1}$, where $M(\xi;\boldsymbol{\theta})$ is the $p \times p$ Fisher information matrix. Therefore the calculation of $M(\xi;\boldsymbol{\theta})$ is an important issue in the construction of D-optimal designs. Other common variance-based optimality criteria also optimise the value of a function of $M(\xi;\boldsymbol{\theta})$. For instance, A-optimal designs minimise trace$[M^{-1}(\xi;\boldsymbol{\theta})]$, which is equivalent to minimising the average asymptotic variance of the parameter estimators. E-optimal designs minimise the variance of the least well estimated normalised contrast $\mathbf{c}^T\boldsymbol{\theta}$, $\mathbf{c} \in \mathbb{R}^p$, $\mathbf{c}^T\mathbf{c} = 1$. Both of these criteria are discussed by Atkinson et al. (2007, Ch. 10). There are also optimality criteria for objectives other than parameter estimation: for example discrimination between models, for which T-optimality may be appropriate (Atkinson et al., 2007, Ch. 20). All of the criteria mentioned above assume that one of the models to be fitted is correct, i.e. is the true data generating process. In general it is possible that all of the models we attempt to fit are incorrect. In this case we say there is *model bias.* For a classic paper which illustrates the potential impact of model bias on the optimal choice of design see Box and Draper (1959).

Parameter dependence. A problem which occurs for models other than the normal-response linear model is that to know which designs are optimal one must know the values of $\boldsymbol{\theta}$. One approach is simply to pick a set of values, $\boldsymbol{\theta}_g$. A D-optimal design, $\xi^*(\boldsymbol{\theta}_g)$, calculated under the assumption that $\boldsymbol{\theta}_g$ is equal to the true parameter vector is referred to as being *locally D-optimal at $\boldsymbol{\theta}_g$*. However, if $\boldsymbol{\theta}_g$ is a poor guess then $\xi^*(\boldsymbol{\theta}_g)$ may be highly inefficient.

Sequential design attempts to overcome this by iteratively calculating optimal designs using information obtained in previous experimental runs. At each iteration, Abdelbasit and Plackett (1983) computed maximum likelihood estimates using the data available so far. With this they calculated a locally optimal one-point design, which was used to obtain the next response. For an example of a Bayesian sequential approach, see Dror and Steinberg (2008). Sequential designs may not be an option if the number of runs is limited, if only a single experiment is possible, or if obtaining the responses is time-consuming. The latter is the case, for instance, in agricultural experimentation where one might have to wait many months between setting up crop varieties and fertilisers, and observing the yield.

Maximin designs are chosen to maximise the worst value,

$$\min_{\boldsymbol{\theta}\in S} \psi(\xi;\boldsymbol{\theta})\,,$$

of a local objective function, $\psi(\xi;\boldsymbol{\theta})$, for $\boldsymbol{\theta}$ in a predefined subset, $S \subseteq \mathbb{R}^p$, of possible parameter values. Example criteria include maximin D-optimality (King and Wong, 2000) and standardised maximin D-optimality (Dette, 1997), with $\psi(\xi;\boldsymbol{\theta}) = \log|M(\xi;\boldsymbol{\theta})|$ and $\psi(\xi;\boldsymbol{\theta}) = \text{eff}(\xi;\boldsymbol{\theta}) = \{|M(\xi;\boldsymbol{\theta})|/\sup_{\xi'}|M(\xi';\boldsymbol{\theta})|\}^{1/p}$ respectively. However, such designs may potentially exhibit poor performance for a large proportion of the parameter space.

In contrast, *Bayesian designs* maximise the value of an objective function of the form

$$\Psi(\xi) = \int_{\mathbb{R}^p} \psi(\xi;\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}\,, \tag{1.2}$$

where $f : \mathbb{R}^p \to [0,\infty)$ is the density function for a prior distribution on $\boldsymbol{\theta}$. Again ψ is a local objective function, though not necessarily the same as before. To obtain a fully principled Bayesian design, $\Psi(\xi)$ should approximate the expected change in (Shannon) information from prior to posterior, as in Chaloner and Verdinelli (1995). A popular implementation of this sets $\psi(\xi;\boldsymbol{\theta}) = \log|M(\xi;\boldsymbol{\theta})|$. In the pseudo-Bayesian approach, it is not assumed that subsequent data analysis will be performed using Bayesian methods. Instead, (1.2) is interpreted as a device which enables us to derive a design whose frequentist performance is good ‘on average’ over the parameter space. In Chapter 2 we adopt a particular pseudo-Bayesian philosophy, assessing the robustness of ξ in terms of the distribution on $\text{eff}(\xi;\boldsymbol{\theta})$ which is induced by the prior. The latter approach originated in Woods et al. (2006).

1.2 Related literature

Our focus throughout Chapters 2–5 is on design for GLMMs and derived models. The class of GLMMs, defined in Section 2.1.1, contains some notable subclasses which are illustrated in Figure 1.1. The important subclasses are the Generalised Linear Models (GLMs), which contain no random effects, and the Linear Mixed Models (LMMs), which contain random effects but have normal response distributions and an identity link function. In addition we have also the Linear Models (LMs), which contain no random effects and have a normal response distribution. In this section we outline the context of our research by presenting an account of the literature

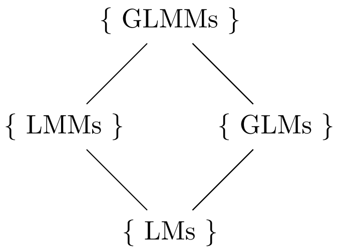

Figure 1.1: Hasse diagram of important subclasses within the family of GLMMs. An upward line from A to B, both sets, indicates that $A \subseteq B$.

on designs for the first three of these model classes. We omit coverage of more basic material on design for LMs. We also refrain from presenting literature on HGLMs since this only relates to Chapter 6.

1.2.1 Design for linear mixed models

Block designs for linear mixed models were studied by Cheng (1995), who was able to prove analytical results of varying generality. A highlight of this paper is a theorem on optimal allocation strategies. This result gives sufficient conditions under which the D-optimal minimally supported design for the LMM takes an appealing form. In particular, given these conditions, the LMM design can be obtained by taking the treatments from a D-optimal LM design, and allocating these to blocks in accordance with a balanced incomplete block design. The paper also gives some numerically calculated locally D-optimal designs for a one-factor quadratic model with two points per block. The method used is not easily adaptable to other examples, but the numerical results provide an interesting comparator for the designs resulting from the theorem. In this specific case, designs resulting from the theorem were extremely close to optimal.

Goos and Vandebroek (2001) developed a point exchange algorithm for the construction of D-optimal designs. The algorithm can easily be extended to different problem structures: the number of factors and block sizes can be varied, and different terms can be included in the predictor. The authors demonstrated that for large values of the correlation parameter, the D-optimal design for a random block effects model is the same as that for a fixed effects model.

Sometimes block designs face additional restrictions. For instance many experiments exhibit a split-plot structure. A split-plot design contains two types of factors: whole-plot and sub-plot factors. Each whole-plot factor is constrained to take the same value for all runs within a given block, whereas the sub-plot factors are free to vary within blocks. In a fixed effects framework, it is impossible to estimate the effect of a whole-plot factor since it is completely confounded with the block effect. An important advantage of random block effects modelling is that it permits estimation of whole-plot factor effects. Jones and Goos (2007) develop methodology for computing optimal split-plot designs under an LMM. A co-ordinate exchange algorithm is used, which eliminates the need for a candidate set of treatments. We consider a similar approach for non-normal responses in Chapter 6. Optimisation methods for further complex design structures have also been developed, for instance split-split-plot designs (Jones and Goos, 2009).

The above papers all address random intercept models only. Ouwens, Tan and Berger (2002) optimised designs for a quadratic model with time as the only factor. Random intercept and random slope terms were also included. The objective was to produce measurement schedules for a growth retardation study. The issue of dependence on the parameters was addressed using a maximin approach. However, the resulting approximate designs were not much more efficient than equally spaced designs of the same size. Berger and Tan (2004) also consider maximin designs for the LMM.

The articles mentioned so far in this section optimise estimation of the fixed effects parameters only, assuming that the variance components are known and that estimation of these is not of interest. For an upcoming article which considers optimisation of the precision of variance component estimators, see Loeza-Serrano and Donev (2012).

1.2.2 Design for generalised linear models

When the response is binary, a logistic regression model is typically used. This is an instance of a GLM. For this model, maximin designs have been discussed by King and Wong (2000) and sequential design by Abdelbasit and Plackett (1983). For count reponses, a Poisson model is often appropriate. Design for this model is considered by Minkin (1993) and Russell, Woods, Lewis and Eccleston (2009).

Much theoretical work on the local design problem for GLMs has focussed on the use of canonical forms to transform the optimisation problem to one in which the objective function does not depend on the parameters. Nonetheless, the resulting transformed design space has constraints which usually do depend on the values of the parameters. Ford, Torsney and Wu (1992) and Atkinson and Haines (1996) discuss canonical forms as well as geometrical methods of finding locally D-optimal designs. A recent advance by Yang, Zhang and Huang (2011) reduces the problem of locally optimal design for a multifactor logistic model to a one-dimensional optimisation problem provided all but one of the covariates are bounded. For the results of the latter to apply, the linear predictor must contain only the first-order effects of the covariates and no interactions.

In the GLM context, various techniques for handling parameter dependence of optimal designs have been explored in depth. An interesting review is given by Khuri, Mukherjee, Sinha and Ghosh (2006). Below we also mention several papers which appeared subsequent to this.

In their paper on Bayesian design, Chaloner and Larntz (1989) applied an expected log-determinant criterion to the logistic model with one factor. This criterion was extended to incorporate uncertainty about the form of the form of the linear predictor and link function by Woods et al. (2006). An algorithmic approach, using simulated annealing, was used to produce designs for multifactor problems, including a four-factor logistic model. Prior to this most papers had concentrated on one or two factors only. Woods et al. (2006) also proposed assessing design performance in terms of the distribution on the local efficiencies induced by the prior distribution. Their methods are able to take into account uncertainty in the form of the predictor, the values of the parameters and also the choice of link function. Gotwalt, Jones and Steinberg (2009) were able to produce more efficient designs for the same problem through the use of a novel numerical integration method.

Dror and Steinberg (2008) combined the sequential and Bayesian approaches by generating a set of candidate augmentation points from the locally optimal design at the posterior median. To discriminate between these candidate points, the objective function of Chaloner and Larntz (1989) was used, with the average being taken instead with respect to the posterior distribution of the parameters. One advantage of this method is that efficient designs can be calculated even when the number of points in the experiment is still small. However, given that small sample size is a concern, it may be better to avoid the use of the information matrix, which is only asymptotically related to the variance of the parameter estimators. One could instead attempt to use alternative estimators whose variance can be computed exactly, such as those considered for logistic regression by Russell, Eccleston, Lewis and Woods (2009).

Dror and Steinberg (2006) obtained robust designs for GLMs by computing large numbers of locally optimal designs and applying k-means clustering to the resulting totality of design points. The designs obtained in this way perform comparably to those of Woods et al. (2006), but are much faster to compute. A further merit is that the locally optimal designs which are required to evaluate the overall performance of the design are already available from the construction algorithm. Russell, Woods, Lewis and Eccleston (2009) adopted a similar strategy, which is particularly effective for the Poisson model due to the availability here of a closed form for the locally optimal designs, thus eliminating the need for any numerical search. However, the formula for the designs applies only to the model with the linear effects of the factors. Moreover, there are mild restrictions on the values of the parameters for which it holds.

Work on GLM designs is ongoing. Yang, Mandal and Majumdar (2012) consider experiments with binary responses and two-level factors. Stufken and Yang (2012) develop theoretical results on optimal designs for binary and count response regression models using fixed effects for blocks.

1.2.3 Generalised linear mixed models

Estimation for GLMMs is significantly more challenging than for LMMs or GLMs because for these models the likelihood contains intractable integrals over the potential values of the random effects (McCulloch and Searle, 2001, Chapter 8). Breslow and Clayton (1993) proposed iterative procedures which approximate the GLMM with LMMs in order to avoid the need to compute difficult integrals, but these approximate procedures were shown to yield biased estimates of the model parameters. It is preferable instead to use the more computationally intensive methods described by McCulloch (1997). These combine an EM or Newton-Raphson algorithm with Metropolis sampling of the posterior distribution of the random effects. However, such methods may potentially be slow. A surprising result due to Lele, Nadeem and Schmuland (2010) makes ML estimation for GLMMs more attractive, interestingly using Bayesian methods but cloning the data so as to filter out the influence of our prior beliefs on the posterior distribution. Truly Bayesian estimation which retains the influence of the prior distribution is also available, for instance see Zeger and Karim (1991) who describe a Gibbs sampling procedure. Bayesian model fitting for GLMMs is implemented in the `R` package `MCMCglmm` by Hadfield (2010). Other aspects of Bayesian data analysis using GLMMs, such as model selection, are also reasonably well developed (e.g. Overstall and Forster, 2010).

Design for GLMMs is complicated by the fact that the information matrix is not available in closed form, and exact evaluation requires computationally intensive numerical integration techniques. Therefore approximations to the information matrix are usually used. These are typically based on the approximate inference techniques of Breslow and Clayton (1993) and Goldstein and Rasbash (1996), and we will follow a similar approach in this thesis. Such approximations have been used by Moerbeek, Van Breukelen and Berger (2001) and Moerbeek and Maas (2005) to produce designs for logistic GLMMs, a problem which was also considered by Ouwens, Tan and Berger (2006). The first two of these design papers gave analytical formulae for the optimal designs which take into account various cost constraints. However, they only considered locally optimal designs. The third design paper uses a maximin criterion and an algorithmic approach to produce robust designs. All three papers are limited by a restriction to dichotomous independent variables, and in the case of the first two only one or two such factors can be considered. With our method, which is based on similar approximations, we shall attempt to deal with multiple continuous factors using a flexible algorithmic approach.

Niaparast (2009) considered design for the log-link Poisson model with random intercept, using a quasi-likelihood approximation to the information matrix. The form of this approximation is simplified by the fact that, in the Poisson case, closed form expressions for the marginal mean and variance are available. This is not true for other distributions or link functions, therefore the approximation can not be adapted to all GLMMs, and the paper did not address the issue of parameter dependence. Niaparast and Schwabe (2013) extend work on quasi-likelihood designs to the Poisson model with a random slope coefficient.

Tekle, Tan and Berger (2008) produced highly efficient maximin designs for binary longitudinal data, using similar information matrix approximations to Moerbeek et al. (2001). The designs consisted of measurement schedules in which all individuals are observed at identical time intervals. The authors consider the optimal number of time points, whereas the corresponding quantity in our work, the block size, is assumed to be fixed by the nature of the experiment. The latter will indeed be the case in many industrial experiments. Whilst various forms of the predictor were considered, the model contains only one independent variable: time.

Sinha and Xu (2011) were able to compute sequential designs for the logistic mixed effects model without resorting to the approximations of Breslow and Clayton (1993), instead using a direct computational approximation to the information matrix. By using an algorithm which adds just one point at a time to the design, the number of possible outcomes is restricted as there are only two possible outcomes per point. This restriction makes the evaluation of the information matrix more feasible. Only locally optimal augmentations were considered.

Ogungbenro and Aarons (2011) considered the use of approximations similar to Breslow and Clayton (1993) for ordinal and count response models. They compared standard errors obtained from simulation with those anticipated by the information matrix approximations. Reasonable agreement was demonstrated.

1.2.4 Further related models

GLMMs are referred to as *conditional models* because the response is assumed to follow a GLM *conditional* upon the realised values of some random effects pertaining to a particular block. Under a conditional model, the responses in the same block are marginally correlated because units in the same block have random effects in common.

An alternative to the conditional model is to assume that any particular response follows a

GLM marginally, but that responses in the same block are correlated. This is referred to instead as marginal modelling. The parameters of this marginal GLM can be estimated using Generalised Estimating Equations (GEEs; Liang and Zeger, 1986), which account for the dependence structure using a working correlation matrix. Liang and Zeger (1986) suggest parameterising the working correlation matrix and attempting to estimate these new correlation parameters from the data. Chaganty and Joe (2004) advocate instead regarding the working correlation matrix as a 'weight matrix' and holding the correlation parameters fixed. This was found to produce more efficient estimates of the marginal model parameters. One concern with marginal modelling is that there may not be any probability model which corresponds to a given correlation structure, as there are bounds on the possible values of the correlation for binary variables (see Chaganty and Joe, 2004, Section 2).

Design for binary data modeled by GEEs has been discussed by Tekle et al. (2008) and Woods and Van de Ven (2011). The former used this approach to take into account autoregressive serial correlations in the longitudinal setting. The latter found that an effective design strategy was to take the design points from a robust unblocked design and allocate these optimally to blocks. When the allocation was chosen to maximise the local objective function at the prior mean, the resulting designs were comparable to the output of an unrestricted numerical search for the optimal Bayesian blocked design. However, the allocation strategy was computationally much cheaper.

Optimal design methodology has also been developed for nonlinear mixed effects models (NLMEs), with particular focus on examples in pharmacokinetics (PK) and toxicokinetics, see for example Mentré, Mallet and Baccar (1997); Gagnon and Leonov (2004). These disciplines study the transport of compounds through the body. Often the compound is a novel drug in an early phase clinical trial. Of primary interest is the time dependence of the concentration of the compound within the bloodstream. Typically parametric compartmental models are used to describe these dynamics. These models are solutions of differential equations approximating the underlying transport mechanism between different 'components' of the body. Random effects are included to model the variation of the parameters between different individuals. In PK experiments, drug concentration measurements are taken on multiple occasions per individual. The design problem is to choose the number of measurements, and the times at which they should be taken. From the perspective of this thesis, there is only one controllable factor (time) in the experiment, and the measurements on an individual patient constitute a block.

There are several commonalities between the optimal design problems in Chapters 2–5, and those for NLMEs. The first is that the information matrix does not have a closed form, necessitating some form of approximation. Another is the parameter dependence of the optimal designs. Retout and Mentré (2003) consider two information matrix approximations. The first linearises the model around the population mean value of the parameters, and is referred to as FO (first order). The second involves a linearisation of the model around a simulated value of the individual parameters. The FO method is similar to one of the approximations we consider, MQL (Breslow and Clayton, 1993). The second method is based on the FOCE (first-order conditional estimation) method of Lindstrom and Bates, which is similar to PQL estimation for GLMMs. The corresponding FOCE information matrix thus is in a similar spirit to our PQL approximation. However, the implementation of the FOCE information matrix involves Monte Carlo integration and is therefore more similar to the PQL approximation of Tekle et al. (2008)

than to ours, since we use a cruder approximation which can be expressed analytically. Atkinson (2008) used Monte Carlo simulations to estimate the marginal mean and covariance of the model. From these quantities it is possible to derive an expression for the information matrix, for details see the reference.

The design approach for NLMEs using FO and FOCE is implemented in the R function PFIM (Bazzoli, Retout and Mentré, 2010). This software allows calculation of optimal designs for many PK models using either the Nelder-Mead simplex algorithm, or a Federov-Wynn algorithm. A graphical user interface is also available to assist the practitioner.

For a cautionary note on the use of linearisation-based approximations to the information matrix in NLMEs, see Mielke and Schwabe (2010).

1.3 Outline of thesis

The generic problem in models with random effects and non-normal response distributions is that the Fisher information matrix involves intractable integrals over the potential values of the random effects. Thus evaluation of $M(\xi;\boldsymbol{\theta})$ is in general a computationally intensive procedure involving numerical integration or Monte Carlo simulation. As a result, direct numerical optimisation of the design is usually much too slow to be practical. Indeed if one computes naïvely the search is likely to take several months.

In Part I, which contains Chapters 2 and 3, we develop design methodology for GLMMs. Several computationally inexpensive, analytical approximations to the information matrix are proposed in Chapter 2. These enable the search time to be vastly reduced. The approximations are based on the approximate estimation procedures, MQL and PQL, and are similar to those used previously in the design literature, e.g. Moerbeek et al. (2001) (for further details, see Section 1.2.3). In Chapter 3, the performance of these approximations is compared through the use of a higher fidelity, more computationally intensive procedure referred to as maximum likelihood by numerical interpolation (MLNI). To our knowledge, no performance comparison between analytical approximations has been attempted before. The MLNI procedure yields locally optimal and Bayesian designs in the case where there are two points per block. A further analytical approximation (AMQL) is proposed in Chapter 3, which yields designs that are almost 100% efficient when compared to the designs from MLNI.

The focus of Part II is the design of dose-response bioassay experiments with heterogeneous individuals. The response is binary, and the event $y = 1$ corresponds to an unrepeatable event, such as death of an individual in the study. Chapter 4 considers the case where we are able to make only one observation per individual. The model in this case is a GLMM, and we explore the performance of some of the approximations of Chapter 2 in this setting. To be able to estimate the parameters of the model with any reasonable amount of precision we need a prior estimate of the degree of heterogeneity among the individuals. Robustness of the estimation to misspecification of this quantity is investigated. In Chapter 5, we consider the case where it is possible to make multiple observations per individual. The model in this case is not a GLMM, owing to fact that the event $y = 1$ is unrepeatable: once an individual 'dies', they cannot yield any further observations. Nonetheless, we can use a computational approximation related to MLNI to derive optimal designs within a restricted class.

Part III contains Chapters 6 and 7. In Chapter 6, we develop design methodology for Hierarchical Generalised Linear Models (HGLMs, Lee and Nelder, 1996). HGLMs constitute an alternative class of models with which it is possible to take into account the two primary features of interest in this thesis, namely (i) non-normal response distributions and (ii) correlation between responses in the same block. For these models, computationally inexpensive asymptotic approximations to the variance-covariance matrix of the parameter estimators are available. An advantage of HGLM design is that consideration of the quality of estimation of the individual random effects is relatively straightforward. In Chapter 7, we consider nonlinear design problems in which the parameter space contains degenerate parameter values, which we refer to as *singularities*. This may lead to the technical difficulty of non-convergence of the objective function when the most common implementation of the Bayesian D-optimality criterion (Chaloner and Verdinelli, 1995) is used. We show, by means of an explicit example, that the issue may arise even when the support of the prior distribution is a bounded interval. Alternative optimality criteria are considered as a solution, in addition to designs with infinite support defined by a probability density function.

Finally, Chapter 8 gives some concluding remarks and suggests potential directions for related future research.

Part I

Designs for Generalised Linear Mixed Models

Chapter 2

Designs from analytical approximations

As discussed in Chapter 1, most standard optimality criteria such as D-, A-, and E- optimality require us to maximise a functional of the Fisher information matrix. Corresponding measures of design performance also involve the information matrix. In Section 2.1 we define the Generalised Linear Mixed Model (GLMM), and in Section 2.2 we explore issues associated with computation of the Fisher information matrix for a GLMM with observations correlated within blocks. The computational cost of the evaluation of the information matrix motivates us to consider computationally cheap analytical approximations which can be used in the search for an optimal design. The GLMM design problem is nonlinear, hence the optimal design depends on the unknown values of the parameters. This parameter dependence is addressed in Section 2.4 through the use of Bayesian designs, and in particular we apply the optimality criteria of Firth and Hinde (1997). Some example designs are computed in Section 2.5. These serve to illustrate some of the principles of block designs in this setting. In Section 2.6, we use our approximations to evaluate designs for the Poisson model, and compare our results with those of Niaparast (2009) and Russell, Woods, Lewis and Eccleston (2009).

2.1 The Generalised Linear Mixed Model

2.1.1 Definition

Suppose that there are q controllable 'treatment' covariates $x_1, \dots, x_q$, each taking values in $[-1, 1]$. Let us denote the response for the jth unit in the ith block by y_{ij}, and the corresponding vector of treatment covariates by $\mathbf{x}_{ij} \in \mathcal{X} = [-1, 1]^q$, for $i = 1, \dots, n$, $j = 1, \dots, m_i$. Let also ζ_i denote the m_i treatment vectors in the ith block, $\zeta_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{im_i}) \in \mathcal{X}^{m_i}$. We say that ζ_i is the exact design for the ith block.

The GLMM can be defined as follows: for each block in the experiment there is an associated vector $\mathbf{u}_i$ of r random effects, conditional upon which the response follows an exponential family distribution,

$$y_{ij}|\mathbf{u}_i \sim \pi(\mu_{ij}),$$

with mean $\mu_{ij} = \mu(\mathbf{x}_{ij}|\mathbf{u}_i)$ and variance $\text{var}(y_{ij}|\mathbf{u}_i) = v(\mathbf{x}_{ij}|\mathbf{u}_i)$. The mean function $\mu(\mathbf{x}|\mathbf{u})$ is defined by

$$\begin{aligned} g(\mu(\mathbf{x}|\mathbf{u})) &= \nu(\mathbf{x}|\mathbf{u};\boldsymbol{\beta}) \\ \nu(\mathbf{x}|\mathbf{u};\boldsymbol{\beta}) &= \mathbf{f}^T(\mathbf{x})\boldsymbol{\beta} + \mathbf{z}^T(\mathbf{x})\mathbf{u} \end{aligned} \tag{2.1}$$

where $\mathbf{f} : \mathcal{X} \to \mathbb{R}^p$ is a known vector of regressor functions, and $\boldsymbol{\beta}$ is a vector containing the p fixed effects parameters. The function $\mathbf{z} : \mathcal{X} \to \mathbb{R}^r$ is also known, and typically will be a subvector of $\mathbf{f}$. The known function g is called the *link function*, and ν in (2.1) is referred to as the *linear predictor*. We also refer to $\eta = \mathbf{f}^T(\mathbf{x})\boldsymbol{\beta}$ as the *fixed part* of the linear predictor. To fully determine the model we must also make assumptions about the distribution of $\mathbf{u}_i$: throughout most of the thesis, we specify $\mathbf{u}_i \sim \text{MVN}(0, G)$ independently for different i, with G a covariance matrix. Often we will also assume that G is known, for more details see Section 2.2. In general, the presence of random effects in the linear predictor introduces a correlation between responses which are in the same block.

The simplest non-trivial random effects structure is exemplified by the random intercept model, in which $r = 1$ so that there is a single, scalar, random effect $u_i \sim N(0, \sigma^2)$ associated with each block. In this case, we have that

$$g(\mu(\mathbf{x}_{ij}|u_i)) = \mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta} + u_i\,, \tag{2.2}$$

and $\mathbf{z}$ is the constant function, $\mathbf{z}(\mathbf{x}) = 1$. If in addition $\mathbf{f} : \mathbf{x} \mapsto (1, \mathbf{x}^T)^T$, then as well as the random intercept the model contains the linear effects of $x_1, \ldots, x_q$.

When the responses are counts, a Poisson distribution may be used for the response together with logarithmic link function, i.e. $g = \log$. In this case we refer to the model (2.1) as a Poisson mixed model.

For binary data, we use a Bernoulli response distribution together with an inverse logistic function as the link, in other words $g(\mu) = \log\{\mu/(1-\mu)\}$. With these additional specifications, we refer to the models (2.1) and (2.2) as the logistic mixed model and logistic random intercept model respectively.

A degenerate case of the logistic random intercept model arises when σ^2 is large. This causes the linear predictor in a block to be swamped by the block effect u_i, which is large in absolute value with high probability, and therefore we have for each $1 \leq i \leq n$ that

$$P(y_{ij} = 1 \,:\, 1 \leq j \leq m_i) \approx 1/2\,,$$

in other words with approximately 50% probability all responses in the ith block are equal to 1. In addition, we have also that

$$P(y_{ij} = 0 \,:\, 1 \leq j \leq m_i) \approx 1/2\,.$$

To see this note that if $\nu_{ij} = \nu(\mathbf{x}_{ij}|\mathbf{u}_i)$ is larger in modulus than a certain magnitude, M_1, then the conditional mean $g^{-1}(\nu_{ij})$ is numerically indistinguishable from one of 0 or 1. For large σ^2, the magnitude, M_2, of u_i required to make $|\eta_{ij}| > M_1$ is negligible compared to the standard deviation of u_i. Therefore the probability that $u_i > M_2$ is essentially 1/2, as is the probability that $u_i < -M_2$. The first of these events leads to a block with all responses equal to 1, and the second to a block with all responses equal to 0, both with high probability. For a more formal proof of these properties using Lebesgue integration theory, see Section 7.9.4.

In these degenerate cases any experimental design will lead to essentially no insight about $\boldsymbol{\beta}$, other than that it is of a different order of magnitude to σ^2. However, if one simply observed data of this type, one might conclude that no dependence on the explanatory variables was present.

2.1.2 Intra-block correlation

In this section we develop some intuition about the strength of dependence between observations in the same block by computing the intra-block correlation as a function of the block effect variance σ^2 in the logistic random intercept model for a particular design.

First of all, we give a description of what happens in the case of an LMM. An analogous linear mixed model to the random intercept model in the previous subsection is

$$y_{ij} = \mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta} + u_i + \epsilon_{ij} \, , \tag{2.3}$$

where the $u_i \sim N(0, \sigma^2)$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ independently. This is a generalised linear mixed model with identity link function and a normal response distribution. In this model, observations in the same block are correlated, specifically

$$\operatorname{corr}(y_{ij}, y_{ik}) = \frac{\sigma^2}{\sigma^2 + \sigma_\epsilon^2} \, , \tag{2.4}$$

for $1 \leq j \neq k \leq m_i$. We refer to (2.4) as the intra-block correlation.

In the logistic random intercept model, in addition to the block effect variance σ^2, the covariance between observations in the same block will depend upon the model parameters and the values of the x_i. There is not a simple closed form expression for this dependence. It is, however, possible to gain insight into how the correlation varies as a function of σ^2 by specifying a design and some values of the parameters.

We consider a two-factor example, $\mathbf{x}_{ij} = (x_1^{(ij)}, x_2^{(ij)})^T$, with linear effects of x_1 and x_2 in the predictor, i.e. $\mathbf{f} : \mathbf{x} \mapsto (1, \mathbf{x}^T)^T$. We set $\boldsymbol{\beta} = (0, 1, 2)^T$ and used the design with a single block ($n = 1$, $m_1 = 4$) given in Table 2.1. Monte Carlo estimates of the mean correlation,

$$\rho = \frac{1}{\binom{m_1}{2}} \left[\sum_{1 \leq j < k \leq m_1} \operatorname{corr}(y_{1j}, y_{1k}) \right] \, ,$$

between observations have been calculated for various σ^2 in the range $[0, 400]$. These were obtained by simulating 1000 possible response patterns for each σ^2 and evaluating the empirical correlation between the responses at the design points. Figure 2.1 shows the correlation for σ^2 up to 50. After this point, the correlation increases much more slowly. Table 2.2 gives rough estimates of the value of σ^2 needed to achieve certain correlations. In this example, the proportion of simulated blocks in which the responses are all identical was approximately equal to the correlation. These results are helpful to inform the choice of prior distribution on σ^2, for example, in Section 2.5.2 the prior mean of $(\beta_0, \beta_1, \beta_2)^T$ is $(0, 1, 2)^T$. At these values, the prior

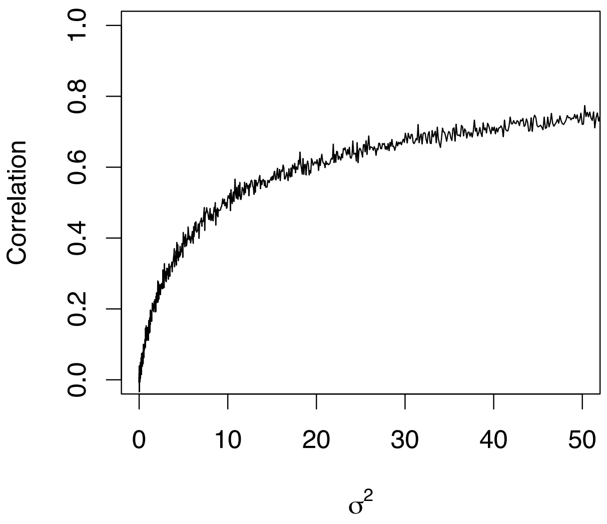

Figure 2.1: Intra-block correlation as a function of σ^2

modal value of $\sigma^2 = 1$ corresponds to a correlation between 0.1 and 0.2. However we must bear in mind that the correlations may not be at all robust to the choice of design, and therefore this study may be used as a very rough guide only.

x_1	x_2
1.00	0.17
-1.00	1.00
-1.00	-0.17
1.00	-1.00

Table 2.1: Design with a single block for the correlation study

2.1.3 Designs

In this chapter the focus is on approximate block designs in which the sizes of all blocks are equal, similar to those considered by Cheng (1995). We explain this further below. For the rest of the chapter we assume that $m_i = m$ for $i = 1, \ldots, n$.

We consider ζ_i and ζ_j, $i \neq j$, to be equivalent (writing $\zeta_i \cong \zeta_j$) if ζ_j can be obtained by rearranging the components of ζ_i, in other words ζ_i and ζ_j contain the same treatments with the same multiplicities. Taking this into account, suppose that there are b distinct ζ_i, $i = 1, \ldots, m$, up to equivalence. Without loss of generality we may reorder the blocks so that ζ_i, $i = 1, \ldots, b$, are distinct (i.e. inequivalent), and for all $j > b$ there is a unique $k \leq b$ with $\zeta_k \cong \zeta_j$. For $k = 1, \ldots, b$, let n_k be the number of blocks using settings ζ_k, in other words n_k is the number of distinct $j \leq n$ such that $\zeta_k \cong \zeta_j$. Writing $w_k = n_k/n$ we have the following concise notation

ρ	σ^2
0.1	0.75
0.2	1.8
0.3	3.5
0.4	6
0.5	10
0.6	20
0.7	50
0.8	100
0.9	340

Table 2.2: Correlation for varying σ^2

for the design, ξ, used:

$$\xi = \left\{ \begin{array}{ccc} \zeta_1 & \cdots & \zeta_b \\ w_1 & \cdots & w_b \end{array} \right\}. \tag{2.5}$$

Clearly for w_k as defined above, nw_k is an integer since $nw_k = n_k$. However, when constructing optimal designs, we search among 'approximate designs' which resemble (2.5) apart from they do not have the restriction that nw_k is an integer. In order to implement an approximate design for a particular finite number of blocks, the weights w_k must effectively be rounded, for details of good rounding procedures see e.g. Pukelsheim and Rieder (1992).

For example, consider a problem in which there are two treatment variables, x_1 and x_2. One might have a design with $b = 2$ support blocks, ζ_1 and ζ_2, each of which has weight 1/2 and contains $m = 2$ treatment vectors. Two potential sets of treatments constituting the blocks ζ_1 and ζ_2 are shown in Figure 2.2. The interpretation of the blocks each having weight 1/2 is that, ideally, $n/2$ of the n blocks in the experiment should use the factor levels in ζ_1 and the other $n/2$ should use those in ζ_2. Of course this will not be possible exactly if n is an odd integer but, provided n is large enough, a close approximation can be obtained by rounding.

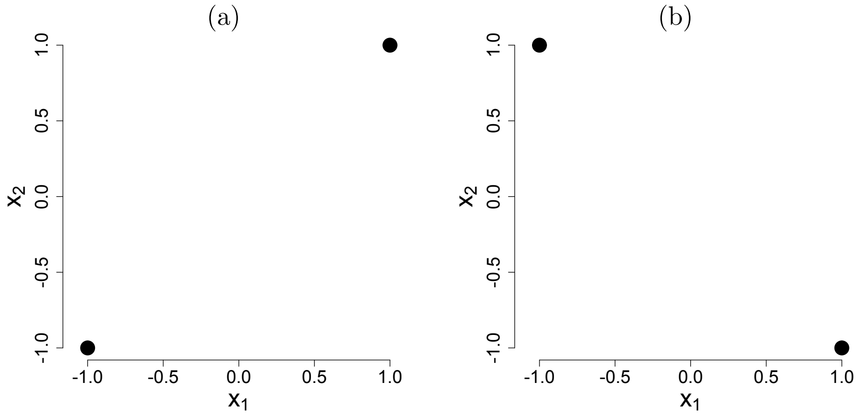

Figure 2.2: Example of an approximate block design consisting of two equally weighted blocks, (a) ζ_1 and (b) ζ_2. Points on the plot correspond to treatment vectors, for instance the bottom-left point in (a) corresponds to $\mathbf{x} = (-1, -1)^T$.

2.2 Information matrix

In this section, we discuss the role of the information matrix, M, and motivate the use of approximations to M for this particular class of models.

Typically optimal designs are chosen to maximise the value of a given functional of the expected Fisher information matrix, M, associated with the estimation problem. For example, under D-optimality one finds the design which maximises the value of $\det(M)$ (Atkinson et al., 2007, p.151). The importance of the information matrix stems from its role in maximum likelihood estimation, where it is proportional to the inverse of the asymptotic variance matrix of the parameter estimators (Davison, 2003, p.118). It can thus be thought of as a measure of the likely precision of the estimators resulting from the experiment.

Let $\boldsymbol{\theta}$ denote the complete set of parameters for the model (2.1). Thus $\boldsymbol{\theta}$ includes the fixed effects parameters $\boldsymbol{\beta}$ as well as parameters specifying the distribution of $\mathbf{u}_i$. Then we shall seek designs which maximise the value of $\det(M_{\boldsymbol{\beta}})$, where $M_{\boldsymbol{\beta}}$ is the information matrix for $\boldsymbol{\beta}$, holding all other components of $\boldsymbol{\theta}$ fixed. The use of $M_{\boldsymbol{\beta}}$ is appropriate when estimating $\boldsymbol{\beta}$ with known variance components. In common with many papers on design for both LMMs (Cheng, 1995; Goos and Vandebroek, 2001) and GLMMs (Moerbeek and Maas, 2005; Tekle et al., 2008; Niaparast, 2009), in this chapter we do not consider the additional variability in $\hat{\boldsymbol{\beta}}$ which is introduced when the variance components also need to be estimated.

For the approximate block design ξ in (2.5), the information matrix $M_{\boldsymbol{\beta}}$ depends on the entire set of parameters $\boldsymbol{\theta}$, and can be decomposed into a weighted sum of information matrices for each support block

$$M_{\boldsymbol{\beta}}(\xi, \boldsymbol{\theta}) = \sum_{k=1}^{b} w_k M_{\boldsymbol{\beta}}(\zeta_k, \boldsymbol{\theta}) \,. \tag{2.6}$$

This follows from the independence of blocks, and is analogous to Cheng (1995, Section 3). The information matrix for the kth block, $k \leq b$, in the design is

$$M_{\boldsymbol{\beta}}(\zeta_k, \boldsymbol{\theta}) = E_{\mathbf{y}_k} \left\{ -\frac{\partial^2 \log p(\mathbf{y}_k | \boldsymbol{\theta}, \zeta_k)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right\} , \tag{2.7}$$

where $\mathbf{y}_k = (y_{k1}, y_{k2}, \ldots, y_{km})^T$ is the vector of responses in the kth block, which is yet to be observed at the planning stage. (Recall that the kth block uses settings ζ_k). The symbol $\frac{\partial}{\partial \boldsymbol{\beta}}$ denotes calculation of the vector of partial derivatives with respect to each of the components of $\boldsymbol{\beta}$. For $\zeta = (\mathbf{x}_1, \ldots, \mathbf{x}_m) \in \mathcal{X}^m$ an arbitrary block, the term $p(\mathbf{y}|\boldsymbol{\theta}, \zeta)$ is the likelihood of the model parameters given the hypothetical vector of responses $\mathbf{y} = (y_1, \ldots, y_m)^T$. For the logistic mixed model with binary response, this is given by

$$p(\mathbf{y}|\boldsymbol{\theta}, \zeta) = \int_{\mathbb{R}^r} \prod_{j=1}^{m} \mu(\mathbf{x}_j|\mathbf{u})^{y_j} \{1 - \mu(\mathbf{x}_j|\mathbf{u})\}^{(1-y_j)} f_{\mathbf{u}}(\mathbf{u}) \, d\mathbf{u} \,, \tag{2.8}$$

where $f_{\mathbf{u}}$ is the density function of a $\mathrm{MVN}(0, G)$ random variable. For the random intercept model (2.8) becomes an integral over $(-\infty, \infty)$ and $f_{\mathbf{u}} = \phi_{\sigma^2}$, where ϕ_{σ^2} is the density function of a $N(0, \sigma^2)$ random variable.

Again for a logistic mixed model with binary response, we can in principle compute the expectations necessary to evaluate (2.6) by considering all possible outcomes in each block. This approach is referred to as *complete enumeration.* Expanding the expectation in (2.7), we have that

$$M_{\boldsymbol{\beta}}(\zeta,\boldsymbol{\theta}) = \sum_{\mathbf{y}\in\{0,1\}^m} \frac{-\partial^2 \log p(\mathbf{y}|\boldsymbol{\theta},\zeta)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\, p(\mathbf{y}|\boldsymbol{\theta},\zeta)\,, \tag{2.9}$$

where the sum is over all possible response patterns in a block. Thus to be able to compute $M(\xi,\boldsymbol{\theta})$ we need only be able to evaluate $p(\mathbf{y}|\boldsymbol{\theta},\zeta)$ and the derivatives of $\log p$ with respect to $\boldsymbol{\beta}$. However, there is in general no closed form expression for the integral in (2.8), therefore we must resort to numerical methods, e.g. quadrature. Moreover there is no closed form for the required derivatives. One approach to calculating the Hessian of $\log p$ is to use finite difference methods of numerical differentiation, combined with quadrature for evaluating integrals. This involves evaluating p at many neighbouring values of $\boldsymbol{\theta}$ for each of the 2^m terms in the sum (2.9). It is clear that this process is computationally involved, and in practice if we evaluate the information matrix in this way the search for an optimal design will likely take several months.

Note that in the case where the response can take infinitely many values, such as when the response is a count with Poisson distribution, the calculation of an expression analogous to (2.9) involves an infinite sum. We must therefore be careful to include sufficiently many terms in order for the partial sum to be close to its limiting value.

Except in the special case of the linear model – see Section 2.2.1 – dependence of the information matrix upon the parameters must be addressed in order to produce a single design for the experiment. For example, one can maximise an objective function which measures the average performance, or expected utility, of a design with respect to a prior distribution on $\boldsymbol{\theta}$. We discuss this in more detail in Section 2.4. However, before doing so, in Section 2.3 we focus on the development of approximate methods for the calculation of $M_{\boldsymbol{\beta}}$ for a given value of $\boldsymbol{\theta}$.

2.2.1 Special cases

To give some context, we now discuss the form of the information matrix in the more straightforward special cases of the model, namely linear models and linear mixed models. We assume for simplicity that $m_i = m$, $i = 1,\ldots,n$.

When the model is linear, with normal response distribution and no random effects, we have

$$M_{\boldsymbol{\beta}}(\xi;\boldsymbol{\theta}) = \sum_{k=1}^{b}\sum_{j=1}^{m} w_k\, \mathbf{f}(\mathbf{x}_{kj})\mathbf{f}^T(\mathbf{x}_{kj})\,.$$

In this case, $M_{\boldsymbol{\beta}}(\xi;\boldsymbol{\theta})$ does not depend on the value of $\boldsymbol{\theta}$ and so optimal designs can be found without prior knowledge about the parameters. With this model, $\mathrm{var}(\hat{\boldsymbol{\beta}}) = (1/n)M_{\boldsymbol{\beta}}^{-1}(\xi,\boldsymbol{\theta})$ for all n. In other words, here the information matrix gives more than just an asymptotic approximation to the estimator variance.

Let us denote by F_i the $m\times p$ model matrix for the ith block of the data, i.e.

$$F_i = \Big[\ \mathbf{f}(\mathbf{x}_{i1})\ \Big|\ \cdots\ \Big|\ \mathbf{f}(\mathbf{x}_{im})\ \Big]^T\,.$$

Furthermore let F be the the full model matrix for the data,

$$F = \left[\; F_1^T \; \middle| \; \cdots \; \middle| \; F_n^T \; \right]^T .$$

Also let $\mathbf{Y} = (y_{11}, y_{12}, \ldots, y_{1m}, y_{21}, \ldots, y_{nm})^T$ be the vector containing the responses written in lexicographical order, and denote by $\mathbf{y}_i = (y_{i1}, \ldots, y_{im})^T$ the vector of responses in the ith block, $1 \leq i \leq n$.

Assume now that we have the random intercept linear mixed model (2.3), with error variance σ_ϵ^2 and random effects variance σ^2. Then the variance matrix of $\mathbf{y}_i$ is

$$\Lambda = \sigma_\epsilon^2 I_m + \sigma^2 \mathbf{1}_m \mathbf{1}_m^T ,$$

where I_m is an $m \times m$ identity matrix and $\mathbf{1}_m$ is an m-vector with all entries equal to 1. The variance of $\mathbf{Y}$ is the block diagonal matrix Π with n repeats of Λ on the diagonal. The maximum likelihood estimator for $\boldsymbol{\beta}$ is

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (1/n) M_{\boldsymbol{\beta}}^{-1} F^T \Pi^{-1} \mathbf{Y} \\ &= (1/n) M_{\boldsymbol{\beta}}^{-1} \sum_{i=1}^{n} F_i^T \Lambda^{-1} \mathbf{y}_i , \end{aligned}$$

where

$$\begin{aligned} M_{\boldsymbol{\beta}}(\xi; \boldsymbol{\theta}) &= (1/n) F^T \Pi^{-1} F \\ &= \sum_{k=1}^{b} w_k F_k^T \Lambda^{-1} F_k . \end{aligned}$$

Equivalent expressions are given by McCulloch and Searle (2001, Section 6.3), but we have modified the notation somewhat. In this model it is also the case that $\mathrm{var}(\hat{\boldsymbol{\beta}}) = (1/n) M_{\boldsymbol{\beta}}^{-1}(\xi; \boldsymbol{\theta})$ for all n, and so the information matrix provides more than just an asymptotic approximation to the estimator variability. Here $\hat{\boldsymbol{\beta}}$ coincides with the generalised least squares estimator (GLS, Draper and Smith, 1998, Section 9.2). GLS will be used in Section 2.3, where we derive approximations to the GLMM information matrix. Note that, given ξ and $\boldsymbol{\theta}$, evaluation of $M_{\boldsymbol{\beta}}(\xi; \boldsymbol{\theta})$ requires only basic operations on relatively small matrices. Thus, evaluation of $M_{\boldsymbol{\beta}}(\xi; \boldsymbol{\theta})$ is computationally inexpensive. Also observe that $M_{\boldsymbol{\beta}}(\xi; \boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$ only through the variance components σ_ϵ^2 and σ^2, and so optimal designs for this model can be derived without knowledge of $\boldsymbol{\beta}$.

2.3 Derivation of approximations

Approximate methods of estimation for GLMMs which avoid the use of quadrature are discussed, among others, by Breslow and Clayton (1993), Rodriguez and Goldman (1995), and Goldstein and Rasbash (1996). These papers propose several approximate procedures, including 'first-order' MQL and PQL together with 'second-order' counterparts, referred to as MQL2 and PQL2 respectively.

In this section, we derive approximations to $\mathrm{var}(\hat{\boldsymbol{\beta}})$ corresponding to these different techniques in order to produce analytically tractable approximations to the information matrix for use in design optimisation. We derive the MQL and PQL approximations to the information matrix, and state the form of an MQL2 approximation which is derived in Section 2.9.

The derivations of the first order methods given by Breslow and Clayton (1993) are comparatively easier to follow, and that paper gives clear expressions for approximations to var($\hat{\boldsymbol{\beta}}$). However, the derivations are not easily extended to higher order approximations. We therefore follow instead the approach of Rodriguez and Goldman (1995) and Goldstein and Rasbash (1996), which revolves around a Taylor series expansion of the inverse link function. The latter two papers are less explicit in terms of expressions for var($\hat{\boldsymbol{\beta}}$).

We recover expressions for MQL and PQL which are identical to those of Breslow and Clayton (1993), and we are also able to derive an expression for MQL2.

The derivations involve several approximation steps which are difficult to justify formally. The ad-hoc nature of these steps is not new to our work, but is already present in the papers mentioned above. In Chapter 3, we compare the resulting approximations in terms of their ability to produce efficient designs.

The expressions given apply to GLMMs with canonical link, and multiple random effects. The MQL2 approximation applies when the random effects are independent, in other words when G is a diagonal matrix.

The MQL and PQL information matrix approximations are similar to the FO and FOCE information matrix approximations proposed for nonlinear mixed effects models by Retout and Mentré (2003). For additional discussion on the commonalities between the design problems here and those for NLMEs, see Section 1.2.4.

2.3.1 Working variates

Goldstein and Rasbash (1996) discussed iterative methods of fitting GLMMs using working variates which are iteratively recalculated and then taken as the response in a weighted least squares regression problem. There are several versions of the methods, which correspond to different forms of the working variate. The different versions are referred to as marginal quasi-likelihood (MQL) and penalised quasi-likelihood (PQL), each of which has a first and second order version. First order MQL and PQL are equivalent to the methods of the same name in Breslow and Clayton (1993). The advantage of using the Goldstein and Rasbash formulation is the extension to second order methods, which are known to be slightly more accurate estimation procedures.

The working variates are functions of (i) the observed data, (ii) some current estimates of the parameter values, and in the case of PQL, (iii) some current predictions of the random effects. In general, the working variate for the first order methods can be written as

$$t_{ij} = \tilde{\nu}_{ij} + \frac{1}{h'(\tilde{\nu}_{ij})} \left[y_{ij} - h(\tilde{\nu}_{ij}) \right] , \tag{2.10}$$

where $\tilde{\nu}_{ij}$ is a current estimate of the linear predictor which depends on the method being used. Above, $h = g^{-1}$ is the inverse link function.

For MQL, we estimate the linear predictor using the current values, $\tilde{\boldsymbol{\beta}}$, of the parameter estimates, and assume that the random effects are approximately zero, so that

$$\tilde{\nu}_{ij}^{\text{MQL}} = \mathbf{f}^T(\mathbf{x}_{ij})\tilde{\boldsymbol{\beta}} = \tilde{\eta}_{ij}\,.$$

In contrast, for PQL we also estimate the values of the random effects using some prediction procedure, e.g. McCulloch and Searle (2001, Ch. 9). For the purposes of deriving the approximations, it is not necessary to consider the details of the procedure used. Thus, the estimate of the linear predictor under PQL is

$$\tilde{\nu}_{ij}^{\text{PQL}} = \mathbf{f}^T(\mathbf{x}_{ij})\tilde{\boldsymbol{\beta}} + \mathbf{z}_{ij}^T\tilde{\mathbf{u}}_i\,,$$

where $\tilde{\mathbf{u}}_i$ is the the current estimate/prediction of the random effects vector $\mathbf{u}_i$.

These working variates can also be shown, using Taylor series expansions of the expected value of the response, to approximately follow a linear model (see Section 2.3.2). Specifically,

$$t_{ij} \approx \nu_{ij} + \frac{1}{h'(\tilde{\nu}_{ij})}\epsilon_{ij}\,, \tag{2.11}$$

where the ϵ_{ij} have mean 0 but are not normally distributed or independent, although we will see in Sections 2.3.4 and 2.3.5 that analytical approximations to the covariance structure of the RHS of (2.11) are available.

The fitting methods proceed by iterating around the following steps until $\tilde{\boldsymbol{\beta}}$ converges:

1. Calculate the values of the working variate, t_{ij}, using the current estimates, $\tilde{\boldsymbol{\beta}}$, of $\boldsymbol{\beta}$ and current predictions, $\tilde{\mathbf{u}}$, of $\mathbf{u}$ if necessary.

2. Regress t_{ij} on the $\mathbf{f}^T(\mathbf{x}_{ij})$ using generalised least squares (GLS, Draper and Smith, 1998, Section 9.2) in order to obtain an updated estimate, $\breve{\boldsymbol{\beta}}$, of $\boldsymbol{\beta}$. To do this we need to know the covariance matrix of the t_{ij}, which is approximated in Sections 2.3.4 and 2.3.5.

3. Set $\tilde{\boldsymbol{\beta}} = \breve{\boldsymbol{\beta}}$.

The limiting value of $\tilde{\boldsymbol{\beta}}$ is used as an overall estimate of $\boldsymbol{\beta}$. In practice, the estimator corresponding to this process may exhibit some bias (for examples see Goldstein and Rasbash, 1996), the magnitude of which depends on the particular version of the method being used. At each stage of the iteration there is a natural information matrix (see Section 2.3.3) for $\boldsymbol{\beta}$ arising from GLS theory, which we use to approximate the true information matrix.

2.3.2 Taylor series expansions

In this section, we use Taylor series expansions to show that the first order working variates approximately follow a linear model.

We proceed by approximating the conditional mean of the response using a Taylor series expansion of the inverse link function, h, around the current estimate of the linear predictor,

$\tilde{\nu}_{ij}$. Specifically we focus on the first order approximation,

$$\begin{aligned} h(\nu_{ij}) &= h(\tilde{\nu}_{ij} + [\nu_{ij} - \tilde{\nu}_{ij}]) \\ &\approx h(\tilde{\nu}_{ij}) + (\nu_{ij} - \tilde{\nu}_{ij})h'(\tilde{\nu}_{ij})\,. \end{aligned} \tag{2.12}$$

We now rewrite the GLMM of Section 2.1 as follows

$$y_{ij} = h(\nu_{ij}) + \epsilon_{ij}\,, \tag{2.13}$$

where the conditional distribution of the error term, ϵ_{ij}, is such that the distribution of y_{ij} remains the same. In particular, ϵ_{ij} satisfies

$$\begin{aligned} E(\epsilon_{ij}|\mathbf{u}_i) &= 0 \\ \operatorname{var}(\epsilon_{ij}|\mathbf{u}_i) &= \operatorname{var}(y_{ij}|\mathbf{u}_i)\,. \end{aligned} \tag{2.14}$$

Note that in general the distribution of ϵ_{ij}, given $\mathbf{u}_i$, will be awkward and not of exponential family form. Substituting (2.12) in (2.13), we obtain a stochastic approximation to y_{ij}, namely

$$y_{ij} \approx h(\tilde{\nu}_{ij}) + (\nu_{ij} - \tilde{\nu}_{ij})h'(\tilde{\nu}_{ij}) + \epsilon_{ij}\,. \tag{2.15}$$

By performing some simple algebraic operations we are able to obtain the form of a working variate which follows a linear model approximately. Subtracting $h(\tilde{\nu}_{ij})$ from both sides of (2.15) we have that

$$y_{ij} - h(\tilde{\nu}_{ij}) \approx (\nu_{ij} - \tilde{\nu}_{ij})h'(\tilde{\nu}_{ij}) + \epsilon_{ij}\,, \tag{2.16}$$

and dividing (2.16) through by $h'(\tilde{\nu}_{ij})$ we find that

$$\frac{1}{h'(\tilde{\nu}_{ij})}\,[y_{ij} - h(\tilde{\nu}_{ij})] \approx \nu_{ij} - \tilde{\nu}_{ij} + \frac{1}{h'(\tilde{\nu}_{ij})}\epsilon_{ij}\,. \tag{2.17}$$

Finally, adding $\tilde{\nu}_{ij}$ to both sides of (2.17), and defining t_{ij} to be equal to the left hand side (thus motivating our earlier definition), we obtain

$$\begin{aligned} t_{ij} &= \tilde{\nu}_{ij} + \frac{1}{h'(\tilde{\nu}_{ij})}\,[y_{ij} - h(\tilde{\nu}_{ij})] \\ &\approx \nu_{ij} + \frac{1}{h'(\tilde{\nu}_{ij})}\epsilon_{ij}\,, \end{aligned} \tag{2.18}$$

as was stated in (2.11). Treating $\tilde{\nu}_{ij}$ as fixed we see using (2.18) that $E(t_{ij}) \approx \nu_{ij}$ and therefore that t_{ij} follows a linear model approximately (albeit with a non-normal response distribution). Clearly the assumption that $\tilde{\nu}_{ij}$ is fixed ignores the fact that the estimates depend on the data.

2.3.3 Relation to information matrix

The approximate information matrices arise out of the iterative methods as follows. For $1 \leq i \leq n$, let $\mathbf{t}_i = (t_{i1}, \ldots, t_{im_i})^T$ be the vector of working variates in the ith block in the data. Denote by V_i the variance matrix of $\mathbf{t}_i$, i.e. the variance of the working variates in the ith block. Finally let V be the block diagonal matrix with blocks V_i, $i = 1, \ldots, n$. Also recall the definitions of the model matrices F_i and F from Section 2.2.1.

Using the approximate model (2.18) the variance of the GLS estimator, $\hat{\boldsymbol{\beta}}$, from step 2 in Section 2.3.1 is approximately

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) \approx \left\{F^T V^{-1} F\right\}^{-1} , \tag{2.19}$$

by generalised least squares theory. The above expression would hold with equality, for all sample sizes, if the linear model approximation (2.18) held exactly. In practice there is a further reason for the approximation symbol in (2.19) since we also approximate $\mathrm{var}(\mathbf{t}_i)$. Equation (2.19) can be rewritten using matrix algebra as

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) \approx \left\{\sum_{i=1}^{n} F_i^T V_i^{-1} F_i\right\}^{-1} ,$$

which can be further rewritten in terms of the b distinct support blocks as

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) \approx \left\{n \sum_{k=1}^{b} w_k F_k^T V_k^{-1} F_k\right\}^{-1} . \tag{2.20}$$

If n is large then we expect that $\mathrm{var}(\hat{\boldsymbol{\beta}}) \approx \mathrm{var}_\infty(\hat{\boldsymbol{\beta}})$, in other words the asymptotic approximation to the variance will be reasonably accurate. Recall that M is related to the asymptotic variance covariance matrix of $\hat{\boldsymbol{\beta}}$ via

$$nM_{\boldsymbol{\beta}}(\xi, \boldsymbol{\theta}) = \mathrm{var}_\infty(\hat{\boldsymbol{\beta}})^{-1} . \tag{2.21}$$

Combining (2.20) and (2.21), we obtain the approximation for the information matrix of the design measure ξ,

$$M_{\boldsymbol{\beta}}(\xi, \boldsymbol{\theta}) \approx \sum_{k=1}^{b} w_k \, F_k^T V_k^{-1} F_k ,$$

Thus, given the definition of a particular working variate, one computes the approximation by evaluating the variance-covariance matrices, V_k, of $\mathbf{t}_k$, $k = 1, \ldots, b$. Different approximations are obtained by considering different forms of working variate.

2.3.4 MQL

In this section we compute the variance-covariance matrix of the t_{ij} under MQL by applying the conditioning identity (McCulloch and Searle, 2001, Ch. 1, p. 11),

$$\mathrm{var}(t_{ij}) = E(\mathrm{var}(t_{ij} \,|\, \mathbf{u}_i)) + \mathrm{var}(E(t_{ij} \,|\, \mathbf{u}_i)) ,$$

to the right hand side of (2.18). We treat the current estimate $\tilde{\nu}_{ij}$ as a fixed quantity. This yields

$$\begin{aligned}\mathrm{var}(t_{ij}) &\approx E\left\{\mathrm{var}\left[\nu_{ij} + \frac{1}{h'(\tilde{\nu}_{ij})}\epsilon_{ij} \,\middle|\, \mathbf{u}_i\right]\right\} + \mathrm{var}\left\{E\left[\nu_{ij} + \frac{1}{h'(\tilde{\nu}_{ij})}\epsilon_{ij} \,\middle|\, \mathbf{u}_i\right]\right\} \\ &= E\left[\frac{1}{h'(\tilde{\nu}_{ij})^2}\mathrm{var}(\epsilon_{ij} \,|\, \mathbf{u}_i)\right] + \mathrm{var}(\nu_{ij}) , \end{aligned} \tag{2.22}$$

where the simplification follows since (i) for the leftmost term, conditional on $\mathbf{u}_i$ the predictor ν_{ij} has no variance and $h'(\tilde{\nu}_{ij})$ is fixed, thus the only term with any variance is ϵ_{ij}, and (ii) for the rightmost term, conditional on $\mathbf{u}_i$, the predictor ν_{ij} is fixed and ϵ_{ij} has zero expectation. Equation (2.22) can be further simplified using (2.14) to

$$\begin{aligned}\operatorname{var}(t_{ij}) &= E\left[\frac{1}{h'(\tilde{\nu}_{ij})^2}\operatorname{var}(y_{ij}\,|\,\mathbf{u}_i)\right] + \mathbf{z}_{ij}^T G \mathbf{z}_{ij} \\ &= E\left[\frac{1}{h'(\tilde{\nu}_{ij})^2} h'(\nu_{ij})\right] + \mathbf{z}_{ij}^T G \mathbf{z}_{ij}\,, \qquad (2.23)\end{aligned}$$

where the second line follows since for the logistic and Poisson models (with log link function) we have that $\operatorname{var}(y_{ij}|\mathbf{u}_i) = h'(\nu_{ij})$. If we approximate both $\tilde{\nu}_{ij}$ and ν_{ij} by $\eta_{ij} = \mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta}$ then (2.23) becomes

$$\operatorname{var}(t_{ij}) = \frac{1}{h'(\eta_{ij})} + \mathbf{z}_{ij}^T G \mathbf{z}_{ij}\,, \qquad (2.24)$$

where the expectation is evaluated trivially because by using $\nu_{ij} \approx \eta_{ij}$ we assume implicitly that the random effects are approximately 0. Using a similar argument we can show that, for $j \neq k$, $\operatorname{cov}(t_{ij}, t_{ik}) = \mathbf{z}_{ij}^T G \mathbf{z}_{ik}$.

2.3.5 PQL

Under PQL we must also condition on the value of $\tilde{\mathbf{u}}_i$. Equation (2.22) becomes

$$\operatorname{var}(t_{ij}) \approx E\left[\frac{1}{h'(\tilde{\nu}_{ij})^2}\operatorname{var}(\epsilon_{ij}\,|\,\mathbf{u}_i, \tilde{\mathbf{u}}_i)\right] + \operatorname{var}(\nu_{ij})\,, \qquad (2.25)$$

which we simplify using the assumption that

$$\operatorname{var}(\epsilon_{ij}|\mathbf{u}_i, \tilde{\mathbf{u}}_i) \approx \operatorname{var}(\epsilon_{ij}|\mathbf{u}_i)\,. \qquad (2.26)$$

We justify this heuristically on the basis that knowing an estimate (i.e. prediction) of $\mathbf{u}_i$ does not give much extra information about $\operatorname{var}(\epsilon_{ij})$ when we already know the value of $\mathbf{u}_i$, except perhaps through indirect information about the fixed effects parameters. This is a simplification and not rigorous mathematics, but it allows us to obtain a form for the expectation which can be evaluated analytically for both the logistic and Poisson models.

Combining (2.25) and (2.26) with the approximation $\tilde{\nu}_{ij} \approx \nu_{ij}$ yields

$$\begin{aligned}\operatorname{var}(t_{ij}) &\approx E\left[\frac{1}{h'(\nu_{ij})^2}\operatorname{var}(\epsilon_{ij}\,|\,\mathbf{u}_i)\right] + \operatorname{var}(\nu_{ij}) \qquad (2.27)\\ &= E\left[\frac{h'(\nu_{ij})}{h'(\nu_{ij})^2}\right] + \operatorname{var}(\nu_{ij}) \qquad (2.28)\\ &= E\left(\frac{1}{h'(\nu_{ij})}\right) + \mathbf{z}_{ij}^T G \mathbf{z}_{ij}\,, \qquad (2.29)\end{aligned}$$

which can be evaluated analytically for the logistic model and Poisson model under log-link as shown in the following. Again we can use a similar argument to show that for $j \neq k$, $\operatorname{cov}(t_{ij}, t_{ik}) = \mathbf{z}_{ij}^T G \mathbf{z}_{ik}$.

Logit link

When the the model is logistic the link function g is the logit (i.e. inverse logistic) function,

$g(\mu) = \log\{\mu/(1-\mu)\}$. We can evaluate the expectation in (2.29) as follows

$$\begin{aligned} E\left(\frac{1}{h'(\nu_{ij})}\right) &= E\left(\frac{1}{\mu_{ij}(1-\mu_{ij})}\right) \\ &= E\left(\frac{(1+e^{\nu_{ij}})^2}{e^{\nu_{ij}}}\right) \\ &= E(e^{-\nu_{ij}} + 2 + e^{\nu_{ij}}) \\ &= 2 + 2e^{\mathbf{z}_{ij}^T G \mathbf{z}_{ij}/2} \cosh(\mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta}) \,, \end{aligned}$$

using that if X is distributed as $N(\mu, \sigma^2)$ then $E(e^X) = e^{\mu+\sigma^2/2}$, together with the fact that $\nu_{ij} \sim N(\mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta}, \mathbf{z}_{ij}^T G \mathbf{z}_{ij})$.

Log link
If a logarithmic link function is used, then the expectation in (2.29) is

$$\begin{aligned} E\left(\frac{1}{h'(\nu_{ij})}\right) &= E\left(\frac{1}{\mu_{ij}}\right) \\ &= E\left(e^{-\nu_{ij}}\right) \\ &= \exp\{-\mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta} + \sigma^2/2\} \,, \end{aligned}$$

using the same result on the expectation of a lognormal random variable.

The PQL approximation we derive in this section is not the same as that used by Tekle et al. (2008) to construct designs for binary time series data. Those authors took as their starting point the approximate PQL variance-covariance matrix of Breslow and Clayton (1993, Section 2.2). The expression for this approximation contains the random effects $\mathbf{u} = (\mathbf{u}_i^T, \ldots, \mathbf{u}_n^T)^T$, which are a (latent) part of the data, and so are not known when designing the experiment. To overcome this problem, the authors used simulated samples of 500 $\mathbf{u}$ vectors from the assumed distribution (fixing the values of the random effects parameters G).

In the above, we do not use simulated $\mathbf{u}$ values. We instead derive a different approximation using a similar approach as for MQL in Section 2.3.4. Our expression can alternatively be obtained by approximating the expectation of the part of the Breslow-Clayton PQL matrix which depends on $\mathbf{u}$, for details see the appendix, Section 2.8. In some sense, our alternative approach attempts to integrate $\mathbf{u}$ out analytically, although to do so we must make a crude approximation. As a consequence, the Monte Carlo PQL approximation of Tekle et al. (2008) is likely to be more accurate than our PQL.

2.3.6 MQL2

In this section we state the form of the approximation for second-order MQL (MQL2). As an estimation method, MQL2 exhibits slightly less bias than MQL (Rodriguez and Goldman, 1995, Section 4.4). Our expression for this approximation only applies when G is a diagonal matrix.

For this approximation, we take a second order expansion of the conditional link function, but omit second order terms in $\boldsymbol{\beta}$ so that we may continue to use linear model theory. More precisely, we use the approximation

$$\begin{aligned} h(\nu_{ij}) &= h\left(\eta_{ij} + \mathbf{f}^T(\mathbf{x}_{ij})(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) + \mathbf{z}_{ij}^T\mathbf{u}_i\right) \\ &\approx h(\eta_{ij}) + h'(\eta_{ij})\left(\mathbf{f}^T(\mathbf{x}_{ij})(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) + \mathbf{z}_{ij}^T\mathbf{u}_i\right) + \frac{h''(\eta_{ij})}{2}(\mathbf{z}_{ij}^T\mathbf{u}_i)^2\,, \end{aligned}$$

and use this to define the working variate

$$\begin{aligned} t_{ij} &= \frac{1}{h'(\eta_{ij})}(y_{ij} - h(\eta_{ij})) + \eta_{ij} - \frac{h''(\eta_{ij})}{2h'(\eta_{ij})}\mathbf{z}_{ij}^T G\mathbf{z}_{ij} \\ &\approx \mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{u}_i + \frac{1}{h'(\eta_{ij})}\epsilon_{ij} + \frac{h''(\eta_{ij})}{2h'(\eta_{ij})}((\mathbf{z}_{ij}^T\mathbf{u}_i)^2 - \mathbf{z}_{ij}^T G\mathbf{z}_{ij})\,. \end{aligned} \tag{2.30}$$

Note that the form of the working variate is essentially the same as that in the first order MQL method, with a simple correction term which ensures that the relation $E(t_{ij}) \approx \mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta}$ is maintained. This correction term is necessary because $(\mathbf{z}_{ij}^T\mathbf{u}_i)^2$ is distributed as $\mathbf{z}_{ij}^T G\mathbf{z}_{ij}$ times a χ_1^2 variable and hence has expectation $\mathbf{z}_{ij}^T G\mathbf{z}_{ij}$.

The derivations of the variance and covariance of the t_{ij} for the second order expansion are much lengthier, and for this reason they are relegated to an appendix, Section 2.9. We give the final expressions here:

$$\begin{aligned} \mathrm{var}(t_{ij}) &= \frac{1}{h'(\eta_{ij})} + \mathbf{z}_{ij}^T G\mathbf{z}_{ij} + (1/2)\left[\frac{h''(\eta_{ij})}{h'(\eta_{ij})}\right]^2(\mathbf{z}_{ij}^T G\mathbf{z}_{ij})^2\,, \\ \mathrm{cov}(t_{ij}, t_{ik}) &= \mathbf{z}_{ij}^T G\mathbf{z}_{ik} + (1/2)\frac{h''(\eta_{ij})}{h'(\eta_{ij})}\frac{h''(\eta_{ik})}{h'(\eta_{ik})}(\mathbf{z}_{ij}^T G\mathbf{z}_{ik})^2\,. \end{aligned}$$

These expressions correspond to those from Section 2.3.4, with additional second order terms in the $\mathbf{z}_{ij}^T G\mathbf{z}_{ik}$. Note that when the link function is logistic, the second derivative satisfies $h''(\eta) = \frac{d}{d\eta}(\mu(1-\mu)) = h'(\eta) - 2h(\eta)h'(\eta) = (1-2\mu)h'(\eta)$, and the above expressions simplify to

$$\begin{aligned} \mathrm{var}(t_{ij}) &= \frac{1}{\mu_{ij}^{(0)}(1-\mu_{ij}^{(0)})} + \mathbf{z}_{ij}^T G\mathbf{z}_{ij} + (1/2)(1-2\mu_{ij}^{(0)})^2(\mathbf{z}_{ij}^T G\mathbf{z}_{ij})^2\,, \\ \mathrm{cov}(t_{ij}, t_{ik}) &= \mathbf{z}_{ij}^T G\mathbf{z}_{ik} + (1/2)(1-2\mu_{ij}^{(0)})(1-2\mu_{ik}^{(0)})(\mathbf{z}_{ij}^T G\mathbf{z}_{ik})^2\,, \end{aligned}$$

where $\mu_{ij}^{(0)} = \mu(\mathbf{x}_{ij}|0)$.

2.4 Robustness of designs to parameter uncertainty

2.4.1 Background and approach

The information matrix for a GLMM, in common with the information matrix for many generalised linear models and nonlinear models, depends on the unknown values of the model parameters. So too, therefore, does the D-optimal design, $\xi_D^* = \arg\max_\xi |M(\xi;\boldsymbol{\theta})|$.

A simple way of obtaining a design when $\boldsymbol{\theta}$ is unknown is to choose a 'guess', $\boldsymbol{\theta}_g$, and use the design which would be optimal were $\boldsymbol{\theta}_g$ in fact correct. This procedure is referred to as *locally optimal design* (Atkinson et al., 2007, Ch. 17). The resulting design, $\xi^*(\boldsymbol{\theta}_g)$, is said to be *locally* *optimal at* $\boldsymbol{\theta_g}$. However, the performance of a design obtained in this way can be hampered by a poor initial estimate. A related concept, which is useful in assessing the performance of a design ξ is that of local efficiency, by which we mean the efficiency of ξ relative to $\xi^*(\boldsymbol{\theta})$ for a particular posited value of $\boldsymbol{\theta}$,

$$\text{eff}(\xi|\boldsymbol{\theta}) = \left\{ \frac{|M(\xi;\boldsymbol{\theta})|}{|M(\xi^*(\boldsymbol{\theta});\boldsymbol{\theta})|} \right\}^{1/p} ,$$

where p is the number of parameters of interest.

In order to construct a design which is robust to potential misspecification of the model parameters, we use a (pseudo-)Bayesian approach, which begins with codifying our prior beliefs about the parameters $\boldsymbol{\theta}$ using a probability distribution, $\mathcal{P}$. It is not however, assumed that the resulting analysis will be Bayesian, or if it is that it will use $\mathcal{P}$. Once we have elicited $\mathcal{P}$, we seek the design which maximises the value of the objective function of Firth and Hinde (1997):

$$I_\alpha(\xi) = \begin{cases} (1/\alpha)\log E_{\boldsymbol{\theta}}\{|M_{\boldsymbol{\beta}}(\xi;\boldsymbol{\theta})|^\alpha\}\,, & \alpha \neq 0\,, \\ E_{\boldsymbol{\theta}}(\log|M_{\boldsymbol{\beta}}(\xi;\boldsymbol{\theta})|)\,, & \alpha = 0\,, \end{cases} \tag{2.31}$$

for some particular choice of α. Those authors show that I_α is a concave function for $\alpha \leq 1/p$, where $p = \dim(\boldsymbol{\beta})$, in which case the criterion satisfies a general equivalence theorem.

A common choice of α in the generalised linear model setting is $\alpha = 0$, for instance Chaloner and Larntz (1989) and Woods et al. (2006). The former authors justify this choice by observing that in this case, maximising I_α is approximately equivalent to maximising the expected posterior gain in Shannon information. Thus the criterion has a fully decision theoretic basis, since it maximises a utility function relating to the anticipated Bayesian analysis. The latter property is less important in the pseudo-Bayesian approach, since we do not assume the analysis will be conducted in a Bayesian fashion.

In some of our work, we use a positive value of $\alpha = 1/p$. An intuitive way of understanding this choice is that we do not want extreme values of the parameters to dominate our design considerations. For logistic GLMMs, uninformative experimental outcomes are highly probable as $\boldsymbol{\theta} \to \infty$ (the case $\sigma^2 \to \infty$ was mentioned in Section 2.1), so it is likely that $|M| \to 0$, and $\log|M| \to -\infty$. Thus it is possible that $\int \log|M(\xi,\boldsymbol{\theta})|f(\boldsymbol{\theta})d\boldsymbol{\theta}$ may fail to converge. We discuss this issue in depth in Chapter 7.

In order to evaluate (2.31), we must use numerical integration methods. Monte Carlo methods or Latin Hypercube Sampling could be employed, but instead in some examples we use the quadrature method of Gotwalt et al. (2009) which has been applied successfully in the case of generalised linear models. In other examples we use a simple discrete approximation to a continuous prior.

Once designs have been obtained, we assess them in terms of their *local efficiency distribution*, which is the distribution induced on $\text{eff}(\xi|\boldsymbol{\theta})$ by the prior distribution on $\boldsymbol{\theta}$. This measure of design robustness has been used previously by Woods et al. (2006) in the context of generalised linear models.

2.5 Examples

2.5.1 Preliminaries

In this section we compute several designs using the approximate methods we have set out so far. We also compare the designs that result from the different approximations, in order to gain an idea of whether the approximations produce consistent answers. First, however we discuss some additional details. We use the integration method of Gotwalt et al. (2009) to calculate the expectation in the Firth-Hinde objective function $I_{1/p}$, defined in (2.31). This technique lends itself most easily to the use of normal prior distributions. Thus we shall adopt the normal as a default prior distribution on the fixed effects parameters. However, the random effects variances must be positive and so for these we use log-normal prior distributions. For details of the implementation, see Appendix 2.10. In future examples we will also use uniform prior distributions on bounded intervals. The objective function in (2.31) is optimised using the transformations of Atkinson et al. (2007, pp. 128–131) which yield an unconstrained optimisation problem. Details of these transformations are also given in Section 3.6. Both the BFGS (Nocedal and Wright, 1999) and Nelder-Mead simplex method (Nelder and Mead, 1965) can be used to perform the optimisation, and both are available through the `R` function `optim` (R Development Core Team, 2012). The BFGS algorithm typically converges in fewer iterations.

To assess the relative performance of the designs found under the different approximations, (i) MQL, (ii) PQL and (iii) MQL2, we must first define measures of local and 'parameter-averaged' efficiency under the different approximations. Using M_a to denote the information matrix under approximation a (one of MQL, PQL or MQL2) we define

$$a\text{-eff}(\xi_1|\xi_2,\boldsymbol{\theta}) = \left\{\frac{|M_a(\xi_1,\boldsymbol{\theta})|}{|M_a(\xi_2,\boldsymbol{\theta})|}\right\}^{1/p}, \tag{2.32}$$

which is the local efficiency (at $\boldsymbol{\theta}$) of design ξ_1 relative to design ξ_2 using approximation a. Let a_1, a_2 and a_3 be MQL, PQL and MQL2 respectively. Also let ξ_i^* be the $I_{1/p}$-optimal design under approximation a_i, $1 \leq i \leq 3$. We define the parameter-averaged efficiency in terms of the objective function $I_{1/p}$ as follows

$$\mathcal{E}(\xi_i^*|\xi_j^*;a_j) = \frac{\exp\{p^{-1}I^{(a_j)}(\xi_i^*)\}}{\exp\{p^{-1}I^{(a_j)}(\xi_j^*)\}} = \frac{\int |M_{a_j}(\xi_i^*,\boldsymbol{\theta})|^{1/p}\,d\mathcal{P}(\boldsymbol{\theta})}{\int |M_{a_j}(\xi_j^*,\boldsymbol{\theta})|^{1/p}\,d\mathcal{P}(\boldsymbol{\theta})}, \tag{2.33}$$

where $1 \leq i,j \leq 3$ and the superscript on the Is denotes that the corresponding approximate information matrix should be used. Note that if the optimisation has correctly converged then (2.33) should be at most 1. In addition observe that if we had that a_j-eff$(\xi_i^*|\xi_j^*,\boldsymbol{\theta}) = e$ for all $\boldsymbol{\theta}$ in the support of $\mathcal{P}$ then we would also have that $\mathcal{E}(\xi_i^*|\xi_j^*;a_j) = e$.

2.5.2 Two factor logistic model

In a two factor model, $\mathbf{x} = (x_1, x_2)^T$. Let the random effects $\mathbf{u}_i = u_i$ be scalar. We assume the following random intercept structure for the linear predictor

$$\nu(\mathbf{x}|u) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u\,. \tag{2.34}$$

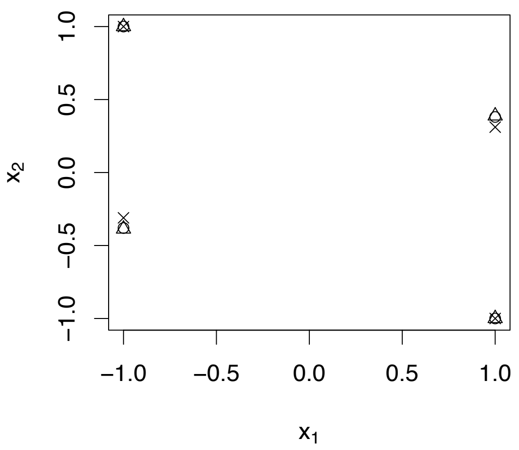

Figure 2.3: Optimal designs for the 2-factor model (2.34) under MQL (△), PQL (∘) and MQL2 (×), using the first prior (common variance of 0.5).

Designs maximising $I_{1/p}$ have been computed for this model with $m = 4$ points per block under two different priors on $\boldsymbol{\theta}$. The priors have a common mean, but the second is more diffuse and thus it is possible to see that the effect of greater uncertainty about the parameters is to increase the number of blocks in the optimal design. This is consistent with the results of Chaloner and Larntz (1989), who found that for logistic models with no block effect the number of treatments in the optimal design increased with the range on uniform priors.

Independent normal priors were used on $(\beta_0, \beta_1, \beta_2, \log \sigma^2)$ with means $(0, 1, 2, 0)$. A common variance of 0.5 was used for the first design, and for the second design the variances on β_0, β_1 and β_2 were increased to 3. Figure 2.3 gives a plot of the single block in the design under the first prior, and Figure 2.4 gives plots of both blocks in the design under the second prior. For each approximation the treatments in the optimal designs are similar, and under prior 2 the allocation to blocks of corresponding points is identical. The weights associated with block (a) in Figure 2.4 were 0.474, 0.475 and 0.476 under MQL, MQL2 and PQL respectively.

Figure 2.5 shows the distributions of the local MQL efficiency, (2.32), for the MQL optimal designs based on a simulation with sample size 1000. This figure was obtained by simulating 1000 parameter vectors from each of the priors, and searching for the locally optimal design at each of those vectors. The resulting designs were then compared to the Bayesian design. The optimisation algorithm used the Bayesian MQL design as an initial design, so that the designs reported as locally optimal were always more efficient than the Bayesian MQL design. Thus the positive density for efficiencies greater than 1 in Figure 2.5 is an artefact of the smoothing method. The mean efficiency under the first prior is 91.1%, under the second (more diffuse) prior it is 72.2%. The lower and upper quartiles of the efficiency under the first prior are 86.0% and 97.7% respectively, whereas under the second prior these are 59.5% and 86.7%. Thus, as one would anticipate, the performance is worse on average and more variable under the more diffuse prior.

A naïve choice might be to use the points from a 2^2 factorial design, in other words $(x_1, x_2) =$

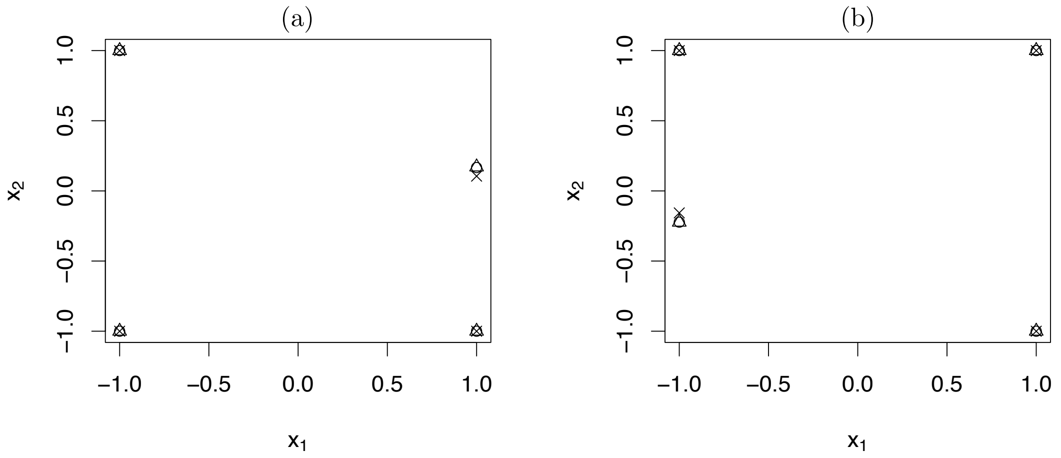

Figure 2.4: (a) First and (b) second blocks of the optimal designs for the 2-factor model (2.34) under MQL ($\triangle$), PQL ($\circ$) and MQL2 ($\times$), using the second prior (common variance of 3).

$(-1,-1)$, $(1,-1)$, $(-1,1)$, and $(1,1)$. Factorial designs are frequently employed when the data can be modelled using a linear model (Atkinson et al., 2007, Ch.7). In this case a single block is large enough to accommodate all points of the factorial design. Figure 2.6 shows the distribution of the local MQL efficiency for the factorial design under the first prior distribution. In this case the factorial performs quite well, and is reasonably robust to different possible values of the parameters. This occurs because the values of the parameters are quite small in this example, so a linear model may reasonably well approximate the logistic model (see Cox, 1988; Woods et al., 2006). For a situation where the factorial (or rather, an allocation of the factorial points) performs less well see Section 3.4.2.

The designs for both priors are compared under each approximation using the objective function efficiency measure (2.33). The results are given in Tables 2.3 and 2.4, and show that each design is highly efficient under all of the different approximations.

	Approximation		
Design	a_1 (MQL)	a_2 (PQL)	a_3 (MQL2)
ξ_1^* (MQL)	1.000000	0.998009	0.999984
ξ_2^* (PQL)	0.998050	1.000000	0.998487
ξ_3^* (MQL2)	0.999983	0.998362	1.000000

Table 2.3: Efficiency of optimal designs under different approximations, 2-factor model, first prior (common variance of 0.5). Explicitly, the entry in the ith row and jth column gives the value of $\mathcal{E}(\xi_i^*|\xi_j^*, a_j)$.

2.5.3 Three factor logistic model

We now have $\mathbf{x} = (x_1, x_2, x_3)^T$. We assume the random effects are scalar, and adopt the following linear predictor

$$\nu(\mathbf{x}|u) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u\,. \tag{2.35}$$

Designs for this model have been computed for a single prior distribution, and the results are presented here. A block size of $m = 3$, which is smaller than the number of fixed parameters,

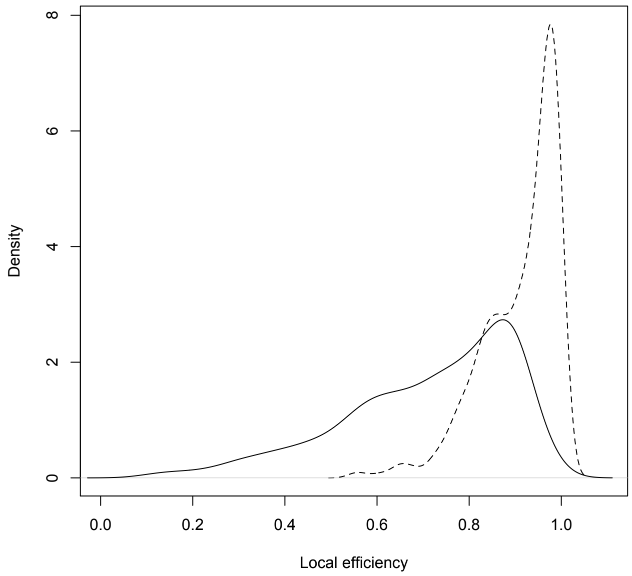

Figure 2.5: Smoothed density estimates of MQL-efficiency distributions for the MQL optimal design under the (a) first prior (broken line), and (b) the second prior (solid line), for the 2-factor model (2.34).

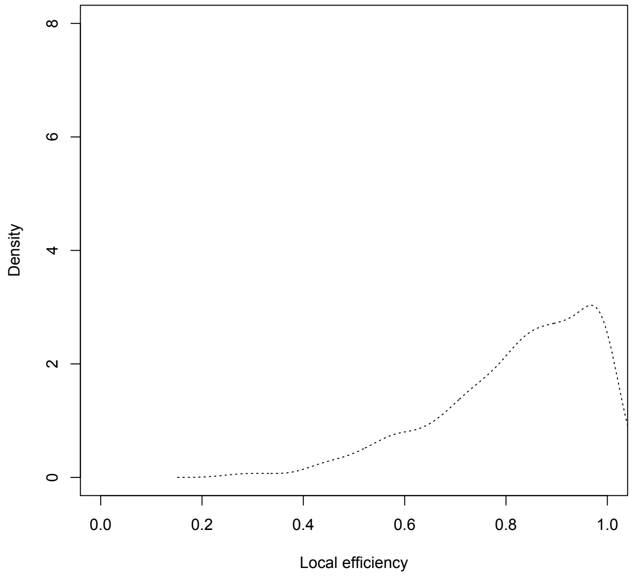

Figure 2.6: Smoothed density estimate of the MQL-efficiency distribution for the factorial design under the first prior, for the 2-factor model (2.34).

	Approximation		
Design	a_1 (MQL)	a_2 (PQL)	a_3 (MQL2)
ξ_1^* (MQL)	1.000000	0.999329	0.999999
ξ_2^* (PQL)	0.999431	1.000000	0.999433
ξ_3^* (MQL2)	0.999999	0.999369	1.000000

Table 2.4: Efficiency of optimal designs under different approximations, 2-factor model, second prior (common variance of 3). Explicitly, the entry in the ith row and jth column gives the value of $\mathcal{E}(\xi_i^*|\xi_j^*, a_j)$.

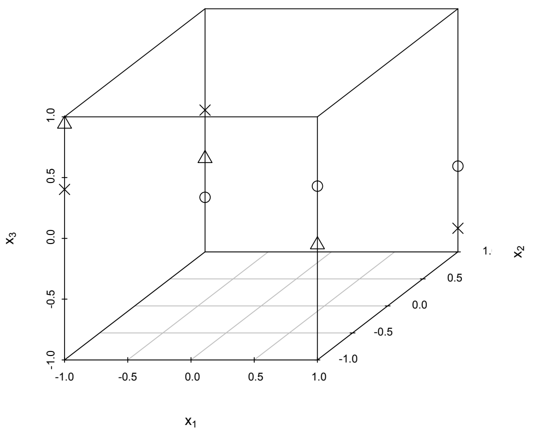

Figure 2.7: Design points for the 3-factor model (2.35) under MQL. Points with the same plotting symbol ($\triangle$, $\circ$ or $\times$) belong to the same block.

was chosen to ensure that the resulting designs would have multiple blocks. Independent normal priors were used on $(\beta_0, \beta_1, \beta_2, \beta_3, \log \sigma^2)$ with means $(0, 1, 2, 5, 0)$ and a common variance of 0.5. In the resulting designs, the values of x_1 and x_2 were close to the endpoints of the interval $[-1, 1]$. The design points, together with their allocation to blocks are shown in Figures 2.7–2.8. The weights of the blocks are given in Table 2.5.

A comparison of the designs under each different approximation, using the objective function measure (2.33) as for the previous example, indicates a small issue of convergence. Namely, the MQL and PQL designs are 0.58% and 0.5% more efficient than the MQL2 design when evaluated under MQL2. Thus in fact it is probably not true that the MQL2 optimal design has only two blocks. Note that these improvements are very small indeed, and all other comparisons yield an efficiency in the range 97.4% to 100%.

Figure 2.9 shows a smoothed density estimate of the local MQL efficiency distribution for the optimal MQL design, based on a simulated sample from the prior of size 1000. Once again, the positive density for efficiencies greater than 1 is an artefact of the smoothing method. The sample estimate of the mean of this distribution is 83.6%, and the median is 85.6%. The lower and upper quartiles are 77.9% and 90.9% respectively. This suggests that the design is reasonably robust to different possible values of the parameter vectors from the prior distribution.

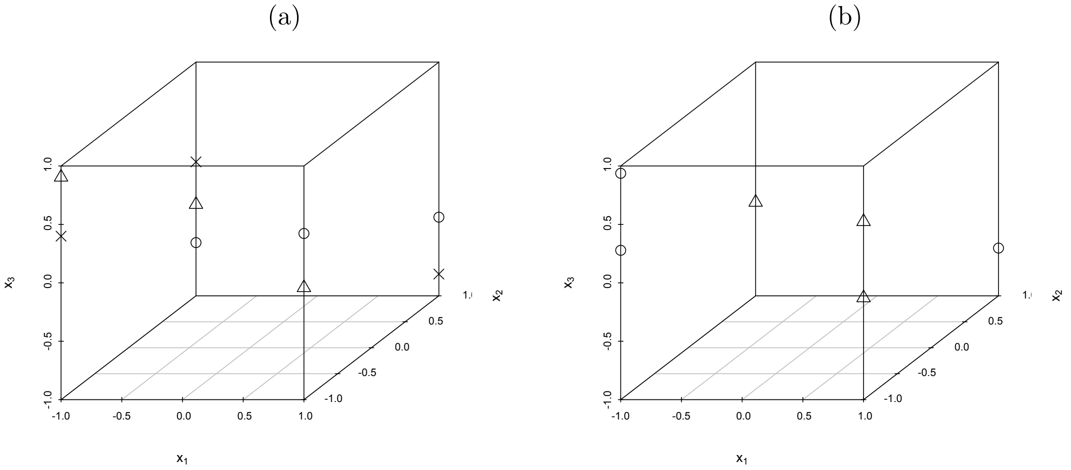

Figure 2.8: Design points for the 3-factor model (2.35) under (a) PQL and (b) MQL2. Points with the same plotting symbol (△, ∘ or ×) belong to the same block.

	Weight		
Block symbol	MQL	PQL	MQL2
∘	.327	.332	.496
△	.294	.302	.504
×	.379	.366	-

Table 2.5: Block weights of designs for the 3-factor model (2.35). Symbols refer to those used in Figures 2.7 and 2.8. The symbol for a PQL block is the symbol used for the corresponding block in the MQL design. There is no correspondence between a block in the PQL/MQL design and one with the same symbol in the MQL2 design.

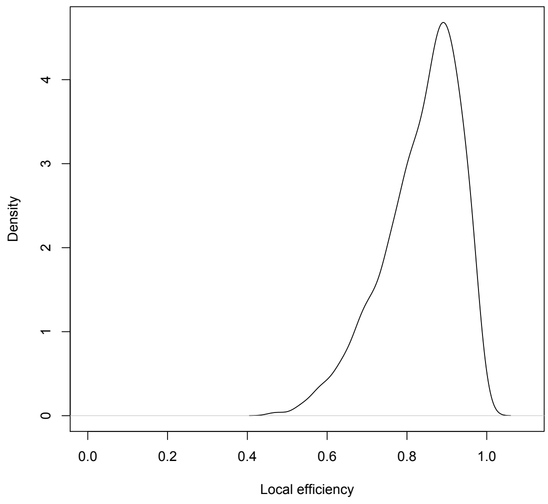

Figure 2.9: Smoothed density estimate of the local MQL-efficiency distribution of the MQL optimal design for the 3-factor model given in (2.35).

2.6 Poisson response

In this section we consider the use of our information matrix approximations to calculate designs for the Poisson model with random intercept. We compare the resulting expressions to those of Niaparast (2009), who considered design for this model using a quasi-likelihood approximation to the information matrix. For more details of the quasi-likelihood approach, see Section 2.6.2. The designs from our methods and those of Niaparast are finally compared to designs for the Poisson model with no random effects, which are obtained using the analytical results of Russell, Woods, Lewis and Eccleston (2009). Throughout this section, the natural logarithm is used as a link function. We largely ignore the issue of parameter dependence of the optimal design by concentrating on locally optimal designs, in other words we assume a value of the parameters.

Niaparast (2009) considered 'doubly approximate' block designs of the following form,

$$\xi = \left\{ \begin{array}{ccc} \zeta_1 & \dots & \zeta_b \\ w_1 & \dots & w_b \end{array} \right\},$$

where the ζ_k, $k = 1, \dots, b$, are themselves approximate designs over $[-1, 1]^q$, i.e.

$$\zeta_k = \left\{ \begin{array}{ccc} \mathbf{x}_{k1} & \dots & \mathbf{x}_{kM_k} \\ \lambda_{k1} & \dots & \lambda_{kM_k} \end{array} \right\}.$$

Above, $w_k, \lambda_{k1}, \dots, \lambda_{kM_k} > 0$, $M_k \geq 1$, $1 \leq k \leq b$, with $\sum_{k=1}^{b} w_k = 1$ and $\sum_{j=1}^{M_k} \lambda_{kj} = 1$. The blocks in this paper were, in theory, repeated measurements on individuals. The author proves that the optimal design of this type treats all individuals identically, in other words the first support block ζ_1 has $w_1 = 1$. This structure is referred to as a single-group design. If we are to apply the same individual design to all participants (i.e. all blocks), then at the very least it needs to be possible to make p observations (runs) per individual (per block) for the fixed effects parameters to be estimable. In general we may not be free to perform so many runs, in which case these designs do not offer much insight.

It is also worth reflecting on the purpose of experimentation. One reason for adopting a random block effects strategy is that we are unable to estimate all of the parameters of interest within a fixed effects framework. Another is that there are many blocks, and it is helpful to introduce some structure to the block effects. One possible setup compatible with a single-group design is an experiment with a small number of blocks and many runs within each block. Here it would be possible to perform intra-block comparisons of all the treatments. The benefits of random-effects modelling are not so clear in such an example, unless one wished to make predictions about future batch effects. However, to be able to make good predictions one is likely to need more blocks.

Our interpretation of these comments is that it is perhaps more realistic and more relevant to restrict the number of support treatments per block. This is the approach we have followed with our definition of a design in Section 2.1.3.

2.6.1 Properties of the model

The Poisson random intercept model uses a log link and satisfies

$$\log E(y_{ij}|u_i) = \nu_{ij} = \mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta} + u_i \,,$$

with conditional variance $\mathrm{var}(y_{ij}|u_i) = E(y_{ij}|u_i) = e^{\mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta}+u_i}$. As before, $\boldsymbol{\beta}$ is the vector of p fixed effects parameters, and the function $\mathbf{f} : [-1,1]^q \to \mathbb{R}^p$ is known. The random effects are independent draws from a $N(0,\sigma^2)$ distribution. For this model, the marginal mean and variance can be computed analytically (Niaparast, 2009). We repeat those results here.

The marginal mean is

$$\bar{\mu}_{ij} = E(y_{ij}) = \exp\{\mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta} + \sigma^2/2\} \,,$$

as e^{u_i} is log-normal. The marginal variance is

$$\begin{aligned} \mathrm{var}(y_{ij}) &= \mathrm{var}\, E(y_{ij}|u_i) + E\,\mathrm{var}(y_{ij}|u_i) \\ &= e^{2\mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta}}\mathrm{var}(e^{u_i}) + e^{\mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta}}E(e^{u_i}) \\ &= e^{2\mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta}}(e^{\sigma^2}-1)e^{\sigma^2} + e^{\mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta}}e^{\sigma^2/2} \\ &= \bar{\mu}_{ij}^2(e^{\sigma^2}-1) + \bar{\mu}_{ij} \,. \end{aligned} \tag{2.36}$$

Responses in different blocks are independent, however responses in the same block have nonzero covariance. For $j \neq k$,

$$\begin{aligned} \mathrm{cov}(y_{ij}, y_{ik}) &= \mathrm{cov}(E(y_{ij}|u_i), E(Y_{ik}|u_i)) + E(\mathrm{cov}(y_{ij}, y_{ik}|u_i)) \\ &= e^{\mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta}}e^{\mathbf{f}^T(\mathbf{x}_{ik})\boldsymbol{\beta}}\mathrm{cov}(e^{u_i}, e^{u_i}) + 0 \\ &= e^{\mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta}}e^{\mathbf{f}^T(\mathbf{x}_{ik})\boldsymbol{\beta}}e^{\sigma^2}(e^{\sigma^2}-1) \\ &= \bar{\mu}_{ij}\bar{\mu}_{ik}(e^{\sigma^2}-1) \,. \end{aligned} \tag{2.37}$$

2.6.2 Quasi-likelihood estimation

The method of quasi-likelihood (Wedderburn, 1974) requires only the mean and variance of the response to be specified, and not a full likelihood. Since the mean and variance of the Poisson random intercept model are analytically tractable, a quasi-likelihood type estimation procedure may be used for this model.

Let us suppose that there are N observations, and the vector of responses, $\mathbf{Y}$, has mean $\boldsymbol{\mu}(\boldsymbol{\beta})$ and variance $\phi^2 V(\boldsymbol{\beta})$, which depends on $\boldsymbol{\beta}$ only through $\boldsymbol{\mu}$, in other words $V = V(\boldsymbol{\mu}(\boldsymbol{\beta}))$. Then the *quasi-score* p-vector is defined as (McCullagh and Nelder, 1989, Chapter 9)

$$U(\boldsymbol{\beta}, \mathbf{Y}) = \phi^2 D^T (V(\boldsymbol{\beta}))^{-1}(\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})) \,,$$

where $D = \frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\beta}^T}$ is the $N \times p$ matrix of partial derivatives of the components of $\boldsymbol{\mu}$ with respect to each component of $\boldsymbol{\beta}$. The quasi-likelihood estimates of $\boldsymbol{\beta}$ are obtained by solving

$$U(\boldsymbol{\beta}, \mathbf{Y}) = \mathbf{0}_p \,, \tag{2.38}$$

where $\mathbf{0}_p = (0, 0, \ldots, 0)^T$ is a p-vector of zeroes. If the observations are independent, the quasi-score vector has an anti-derivative, in other words there exists a scalar function $Q(\boldsymbol{\beta}, \mathbf{Y})$ such that

$$\frac{\partial Q}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}, \mathbf{Y}) = U(\boldsymbol{\beta}, \mathbf{Y}) \,.$$

In this case Q is referred to as the *quasi-likelihood* function. Moreover in this case, solving (2.38) is equivalent to finding the estimates of $\boldsymbol{\beta}$ which maximise Q.

Subject to some regularity conditions, the estimators obtained from solving (2.38) are asymptotically normally distributed with variance-covariance matrix equal to the inverse of the variance of the quasi-score

$$M_{\mathrm{QL}}(\boldsymbol{\beta}) = D^T (V(\boldsymbol{\beta})^{-1}) D \,. \tag{2.39}$$

Thus M corresponds to the information matrix under likelihood theory. We refer to M as the quasi-likelihood information matrix.

However, when we have dependent data there is no guarantee that a proper quasi-likelihood function exists (McCullagh and Nelder, 1989). There is some dispute as to the correct interpretation of quasi-likelihood in this case, however we may still regard (2.38) as a set of estimating equations. Generalised estimating equations (GEEs; Liang and Zeger, 1986) extend these estimating equations to the case of dependent data, and also allow the variance-covariance structure to be incorrectly specified. Moreover the above paper shows that asymptotic results hold even when the variance specification is wrong, and that (2.39) is correct when the variance is correctly specified. Thus we can regard the quasi-score estimating equations for dependent data as an instance of GEEs when the variance *is* in fact correctly specified.

Niaparast (2009) defines D-optimal designs for the Poisson random intercept model to be those which maximise $\det(M_{\mathrm{QL}})$. We can compute M_{QL} using the formula (2.39) together with the analytical expressions for the components of V, equations (2.36) and (2.37).

2.6.3 Designs

We computed locally D-optimal designs for the Poisson mixed model by numerically optimising the determinants of the MQL, PQL and QL approximate information matrices. This was done for the two-factor model with linear effects plus random intercept; in other words the model with linear predictor

$$\nu(\mathbf{x}|u) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \,.$$

The values assumed for the fixed effects parameters were $(\beta_0, \beta_1, \beta_2) = (0, 1, 2)$, and a range of values of the block effect variance σ^2 was used.

Russell, Woods, Lewis and Eccleston (2009) derived an analytical form for the D-optimal approximate design in the case of the first order Poisson model with no random effects and no blocking. For the model with two explanatory variables, this optimal design has 3 equally weighted support points. More specifically, when we take $\boldsymbol{\beta} = (0, 1, 2)^T$ as above, the design measure on $\mathcal{X} = [-1, 1]^2$ is

$$\xi = \left\{ \begin{array}{ccc} (-1, 1) & (1, 0) & (1, 1) \\ 1/3 & 1/3 & 1/3 \end{array} \right\} .$$

σ^2	Method	Design		
.16	quasi	$(1,1)$	$(-1,1)$	$(1,-0.092)$
	PQL			$(1,-0.087)$
	MQL			$(1,-0.089)$
.81	quasi			$(1,-0.096)$
	PQL			$(1,-0.092)$
	MQL			$(1,-0.094)$
2	quasi			$(1,-0.098)$
	PQL			$(1,-0.093)$
	MQL			$(1,-0.096)$
10	quasi			$(1,-0.098)$
	PQL			$(1,-0.064)$
	MQL			$(1,-0.098)$

Table 2.6: Optimal designs for Poisson mixed model, 3 points per block. All are single block designs and contain the points $(1,1)$ and $(-1,1)$.

In order to be able to use the above analytical result as a point of reference, we fixed a block size of 3 for the examples, and considered the value $\sigma^2 = 0$. In this case, the designs found using numerical optimisation contained the points anticipated by the theory. Table 2.6.3 shows the resulting experimental designs under the different approximations for each of four values of $\sigma^2 > 0$. In each case, the design contained a single block, in other words $b = 1$, $w_1 = 1$. For the smaller values of the block variance, the designs from the different approximations appear quite similar. However, they clearly differ from the designs obtained not taking the block effect into account. When $\sigma^2 = 10$, the MQL and QL designs are close together, but the PQL design is a little different.

2.7 Discussion

In this chapter we have proposed methods of obtaining approximate Bayesian designs for GLMMs where the observations are correlated within blocks, using approximations to the information matrix and the criterion of Firth and Hinde (1997). Designs have been calculated for a few specific forms of the linear predictor, and a number of prior distributions. We have successfully replicated the result that increasing prior vagueness leads to more support points, which in this case means blocks, in the optimal design. Moreover, we have seen that for the values of σ^2 in this chapter, the different approximations yield fairly consistent designs in terms of efficiency. However, we have not yet assessed the impact of this work by comparing the efficiency of designs obtained to those from simpler methods, or indeed to designs obtained using the exact information matrix.

In Chapter 3 we shall compare the different approximations by using various benchmark problems, and we find some evidence that, when the value of σ^2 is large enough for the different approximations to diverge, MQL outperforms PQL and MQL2. This agrees with the results for the one-factor model in Chapter 4, although in this special case designs based on a simple approximation to the marginal mean are able to do much better than either MQL or PQL.

Despite the difficulty of obtaining maximum likelihood designs in the general case, we are able in Chapter 3 to develop a methodology to do so when the block size is two. Moreover the approach is such that Bayesian designs can be computed, and local efficiencies can be compared across the optimal designs from each of the different methods.

The approximations of this chapter should be readily adaptable to more models than the random intercept model. It would be interesting to attempt to compute designs for models including block-variable interactions, where the effects of the x_i are allowed to vary randomly across the blocks. Examples of applications with such a structure in the linear predictor include split-plot experiments in film manufacturing (Robinson et al., 2004), and semiconductor manufacture (Robinson et al., 2006), although in these cases the response is neither Bernoulli nor Poisson but Gamma distributed. Our approximations should however also carry over to this case, we must simply substitute the correct link function in equations (2.24) and (2.29).

2.8 Appendix: Alternative derivation of PQL

In this Section, we provide an alternative derivation of the PQL approximation given in Section 2.3.5. The starting point is the PQL variance expression of Breslow and Clayton (1993), which depends on $\mathbf{u}$. We wish to take the expectation with respect to $\mathbf{u}$, but to do so we must make a further approximation.

From Breslow and Clayton (1993, Section 2.2), the approximate PQL variance-covariance matrix for the parameter estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is (translated into our notation),

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) \approx (F^T V^{-1} F)^{-1} \, . \tag{2.40}$$

The matrix V is block diagonal with blocks V_i, $i = 1, \ldots, n$, given by

$$\begin{aligned} V_i &= \mathcal{V}(\zeta_i|\mathbf{u_i}) \\ &= W(\zeta_i|\mathbf{u}_i)^{-1} + Z(\zeta_i) G Z(\zeta_i)^T \, , \end{aligned}$$

where for an arbitrary block, $\zeta = (\mathbf{x}_1, \ldots \mathbf{x}_m)$,

$$Z(\zeta) = [\mathbf{z}(\mathbf{x}_1) \, \ldots \, \mathbf{z}(\mathbf{x}_m)]^T \, ,$$

and $W(\zeta|\mathbf{u})$ is the diagonal matrix of conditional variances of the responses given the random effects $\mathbf{u}$, in other words

$$\begin{aligned} W(\zeta|\mathbf{u}) &= \mathrm{diag}\left\{ v(\mathbf{x}_i|\mathbf{u}) \, : \, 1 \le j \le m \right\} \\ &= \mathrm{diag}\left\{ h'(\eta_i) \, : \, 1 \le j \le m \right\} \, , \end{aligned}$$

with $\eta_i = \mathbf{f}^T(\mathbf{x}_i)$, $1 \le i \le m$. Thus, the PQL variance approximation satisfies

$$\begin{aligned} \frac{1}{n} \mathrm{var}_{\mathrm{PQL}}(\hat{\boldsymbol{\beta}})^{-1} &= \frac{1}{n} \sum_{i=1}^{n} F_i^T \mathcal{V}^{-1}(\zeta_i|\mathbf{u_i}) F_i \\ &= \sum_{k=1}^{b} \frac{n_k}{n} F_k^T \left\{ \frac{1}{n_k} \sum_{\{i \,:\, \zeta_i \cong \zeta_k\}} \mathcal{V}^{-1}(\zeta_k|\mathbf{u_i}) \right\} F_k \, , \end{aligned}$$

$$\to \sum_{k=1}^{b} w_k F_k^T E_{\mathbf{u}}\{\mathcal{V}^{-1}(\zeta_k|\mathbf{u})\} F_k \qquad \text{almost surely as } n \to \infty\,.$$

The second line can be obtained using the following idea. Since we have ordered the blocks such that $\zeta_1, \dots, \zeta_b$ are the distinct blocks among $\zeta_1, \dots, \zeta_n$, we may group terms of the sum whose blocks ζ_i are equivalent. The third line follows using the strong law of large numbers as follows. As $n \to \infty$, so too $n_k = w_k n \to \infty$. The terms in the inner sum, $\mathcal{V}^{-1}(\zeta_k|\mathbf{u}_i)$, $i : \zeta_i \cong \zeta_k$, form an IID sample of size n_k. The mean of this sample converges almost surely to the population mean.

No closed form for $E_{\mathbf{u}}(\mathcal{V}^{-1}(\zeta|\mathbf{u}))$ exists. However, as we saw in Section 2.3.5, $E(W^{-1}(\zeta|\mathbf{u}))$ does have an analytical form for both the logistic and Poisson model. We now make the following approximation, making no claim that it is close,

$$\begin{aligned} E_{\mathbf{u}}(\mathcal{V}^{-1}(\zeta_k|\mathbf{u})) &= E\{(W(\zeta_k|\mathbf{u})^{-1} + Z(\zeta_k) G Z(\zeta_i)^T)^{-1}\} \\ &\approx (E\{W(\zeta_k|\mathbf{u})^{-1}\} + Z(\zeta_k) G Z(\zeta_k)^T)^{-1} \\ &=: \bar{V}_k^{-1}\,. \end{aligned}$$

This yields the PQL approximation to the information matrix,

$$\begin{aligned} M(\xi, \boldsymbol{\theta}) &= \sum_{k=1}^{b} w_k\, F_k \bar{V}_k^{-1} F_k^T \\ &= \sum_{k=1}^{b} w_k\, F_k \left\{E\{W(\zeta_k|\mathbf{u})^{-1}\} + Z(\zeta_k) G Z(\zeta_k)^T\right\}^{-1} F_k^T\,, \end{aligned}$$

which is identical to the expression we obtained previously.

2.9 Appendix: Derivation of the MQL2 approximation

2.9.1 Variance of the working variate

To compute the variance of t_{ij} we condition on $\mathbf{u}_i$. Thus, using the approximate model (2.30),

$$\begin{aligned} \operatorname{var}(t_{ij}) &= E(\operatorname{var}(t_{ij} \mid \mathbf{u}_i)) + \operatorname{var}(E(t_{ij} \mid \mathbf{u}_i)) \\ &\approx E\left(\frac{h'(\nu_{ij})}{h'(\eta_{ij})^2}\right) + \operatorname{var}\left[\mathbf{z}_{ij}^T \mathbf{u}_i + \frac{h''(\eta_{ij})}{2h'(\eta_{ij})}(\mathbf{z}_{ij}^T \mathbf{u}_i)^2\right]\,. \end{aligned}$$

We may approximate the first term in the same manner as for first order MQL. The second term may be evaluated by comparison to a non-central χ_1^2 distribution. Note that for v an arbitrary random variable and h'', h' deterministic functions, by completing the square,

$$\begin{aligned} \operatorname{var}\left[v + \frac{h''}{2h'} v^2\right] &= \left(\frac{h''}{2h'}\right)^2 \operatorname{var}\left[v^2 + \frac{2h'}{h''} v\right] \\ &= \left(\frac{h''}{2h'}\right)^2 \operatorname{var}\left[\left(v + \frac{h'}{h''}\right)^2\right]\,. \end{aligned}$$

In Secton 2.9.2 we show that when $v \sim N(0,\tau^2)$ and λ is non-random we have that $\mathrm{var}\{(v+\lambda)^2\} = 2\tau^2(\tau^2+2\lambda^2)$. Therefore, letting $\lambda = \frac{h'}{h''}$, and $v = \mathbf{z}_{ij}^T\mathbf{u}_i$ the above becomes

$$\begin{aligned}\frac{h''^2}{4h'^2}\,2\,\mathrm{var}(\mathbf{z}_{ij}^T\mathbf{u}_i)\left\{\mathrm{var}(\mathbf{z}_{ij}^T\mathbf{u}_i)+2\left(\frac{h'}{h''}\right)^2\right\} &= \frac{\mathrm{var}(\mathbf{z}_{ij}^T\mathbf{u}_i)^2}{2}\left(\frac{h''}{h'}\right)^2+\mathrm{var}(\mathbf{z}_{ij}^T\mathbf{u}_i)\\ &= (1/2)^2\left(\frac{h''(\eta_{ij})}{h'(\eta_{ij})}\right)^2(\mathbf{z}_{ij}^TG\mathbf{z}_{ij})^2+\mathbf{z}_{ij}^TG\mathbf{z}_{ij}\,.\end{aligned}$$

Therefore the overall expression for the variance is

$$\mathrm{var}(t_{ij}) = \frac{1}{h'(\eta_{ij})}+\mathbf{z}_{ij}^TG\mathbf{z}_{ij}+(1/2)\left(\frac{h''(\eta_{ij})}{h'(\eta_{ij})}\right)^2(\mathbf{z}_{ij}^TG\mathbf{z}_{ij})^2\,.$$

2.9.2 Variance of a non-central χ^2 distribution.

In fact we compute the variance of $v = (u+\lambda)^2$ where u is $N(0,\tau^2)$ and λ is non-random. First of all we note that the mean is

$$E(v) = E(u^2+2\lambda u+\lambda^2) = \tau^2+\lambda^2\,.$$

Now we apply the usual formula for the variance,

$$\begin{aligned}\mathrm{var}(v) &= E(v^2)-(E(v))^2\\ &= E((u+\lambda)^4)-(\tau^2+\lambda^2)^2\\ &= E(u^4+4u^3\lambda+6u^2\lambda^2+4u\lambda^3+\lambda^4)-(\tau^2+\lambda^2)^2\\ &= 3\tau^4+6\tau^2\lambda^2+\lambda^4-(\tau^2+\lambda^2)^2\\ &= 2\tau^4+4\tau^2\lambda^2\,,\end{aligned}$$

where we computed $E(u^4)=3\tau^4$ and $E(u^3)=0$ using moment generating functions.

2.9.3 Covariance of the working variates

In the above we neglected computing the covariances of the MQL2 working variate at different points in the same block. This is an essential part of the approximation, therefore we present this calculation here. For $1 \le i \le n$, $1 \le j \ne k \le m_i$, we condition on the random effect vector $\mathbf{u}_i$ to obtain

$$\begin{aligned}\mathrm{cov}(t_{ij},t_{ik}) &= \mathrm{cov}(E(t_{ij}|\mathbf{u}_i),E(t_{ik}|\mathbf{u}_i))+E(\mathrm{cov}(t_{ij},t_{ik}|\mathbf{u}_i))\\ &= \mathrm{cov}\left\{\left(\mathbf{z}_{ij}^T\mathbf{u}_i+\frac{h''(\eta_{ij})}{2h'(\eta_{ij})}(\mathbf{z}_{ij}^T\mathbf{u}_i)^2\right),\left(\mathbf{z}_{ik}^T\mathbf{u}_i+\frac{h''(\eta_{ik})}{2h'(\eta_{ik})}(\mathbf{z}_{ik}^T\mathbf{u}_i)^2\right)\right\}\,,\end{aligned}$$

where $\eta_{ij} = \mathbf{x}_{ij}\tilde{\boldsymbol{\beta}}$ is the current estimate of the fixed part of the linear predictor neglecting random effects. The second line above follows since $\mathrm{cov}(t_{ij},t_{ik}|\mathbf{u}_i) = 0$, due to conditional independence of t_{ij} and t_{ik}. Defining

$$U_j' = \mathbf{z}_{ij}^T\mathbf{u}_i+\frac{h''(\eta_{ij})}{2h'(\eta_{ij})}(\mathbf{z}_{ij}^T\mathbf{u}_i)^2\,,$$

we have that

$$\operatorname{cov}(t_{ij}, t_{ik}) = E(U'_j U'_k) - E(U'_j)E(U'_k)\,. \tag{2.41}$$

Multiplying out U'_j and U'_k directly gives the following expression for the product,

$$\begin{aligned} U'_j U'_k &= (\mathbf{z}_{ij}^T\mathbf{u}_i)(\mathbf{z}_{ik}^T\mathbf{u}_i) + \frac{h''(\eta_{ij})}{2h'(\eta_{ij})}(\mathbf{z}_{ij}^T\mathbf{u}_i)^2(\mathbf{z}_{ik}^T\mathbf{u}_i) \\ &\quad + \frac{h''(\eta_{ik})}{2h'(\eta_{ik})}(\mathbf{z}_{ij}^T\mathbf{u}_i)(\mathbf{z}_{ik}^T\mathbf{u}_i)^2 + \frac{h''(\eta_{ij})h''(\eta_{ik})}{4h'(\eta_{ij})h'(\eta_{ik})}(\mathbf{z}_{ij}^T\mathbf{u}_i)^2(\mathbf{z}_{ik}^T\mathbf{u}_i)^2\,, \end{aligned} \tag{2.42}$$

whose terms are polynomials in the components of the $\mathbf{u}_i$. The degrees of the polynomial from the terms 1, 2, 3 and 4 are 2, 3, 3 and 4 respectively. We now look at these polynomials in more detail to compute the expectation $E(U'_jU'_k)$, and eventually the covariance $\operatorname{cov}(t_{ij}, t_{ik})$.

We now make use of the assumption, made for MQL2 only, that G is a diagonal matrix. We aim to show that the second and third terms of (2.42) contribute nothing to the expectation $E(U'_jU'_k)$. Let us denote the jth component of $\mathbf{u}_i$ by u_{ij}, $j = 1, \ldots, r$. Then the second and third terms of (2.42) are both linear combinations of third order terms in the u_{ij}. Since G is diagonal, different components of $\mathbf{u}_i$ are independent, and so

$$\begin{aligned} & E(u_{il}^2 u_{im}) = E(u_{il}^2)E(u_{im}) = 0\,, \\ \text{and} \quad & E(u_{il}u_{im}u_{in}) = E(u_{il})E(u_{im})E(u_{in}) = 0\,, \end{aligned}$$

where l, m and n are arbitrary distinct indices. Therefore all third order terms in the u_{ij} vanish when we take the expectation.

We may obtain an expression for the expectation of the first term in (2.42) by writing out the linear combination $\mathbf{z}_{ij}^T\mathbf{u}_i$ explicitly,

$$\begin{aligned} E\left\{(\mathbf{z}_{ij}^T\mathbf{u}_i)(\mathbf{z}_{ik}^T\mathbf{u}_i)\right\} &= \sum_{1\le l,m\le r} z_{ij}^{(l)} z_{ik}^{(m)} E(u_{il}u_{im}) \\ &= \sum_{l=1}^{r} z_{ij}^{(l)} z_{ik}^{(l)} E(u_{il}^2) \\ &= \sum_{l=1}^{r} z_{ij}^{(l)} z_{ik}^{(l)} G_{ll} \\ &= \mathbf{z}_{ij}^T G \mathbf{z}_{ik}\,, \end{aligned}$$

where $z_{ij}^{(l)}$ denotes the lth co-ordinate of the vector $\mathbf{z}_{ij}$. The second and third lines follow since if $k \neq l$ then $E(u_{ik}u_{il}) = E(u_{ik})E(u_{il}) = 0$ by independence, and also $E(u_{il}^2) = \operatorname{var}(u_{il}) = G_{ll}$ as u_{il}^2 follows a $G_{ll}\chi_1^2$ distribution.

We now consider the expectation of the fourth term in (2.42), for the moment forgetting the factor involving derivatives of f. By writing out the linear combinations explicitly we obtain

$$E\left\{(\mathbf{z}_{ij}^T\mathbf{u}_i)^2(\mathbf{z}_{ik}^T\mathbf{u}_i)^2\right\} = \sum_{1\le s,t,v,w\le r} z_{ij}^{(s)} z_{ij}^{(t)} z_{ik}^{(v)} z_{ik}^{(w)} E(u_{is}u_{it}u_{iv}u_{iw})\,. \tag{2.43}$$

We now partition the set of 4-tuples $\mathcal{S} = \{(s,t,v,w) : 1 \le s,t,v,w \le r\}$ according to the number, and multiplicities, of distinct values taken by s,t,v,w. Any (s,t,v,w) must belong to one of $\mathcal{S}_1, \ldots, \mathcal{S}_5 \subset \mathcal{S}$ defined as follows:

1. $\mathcal{S}_1 = \{(s,s,s,s) : 1 \leq s \leq r\}$, i.e. indices all equal
2. $\mathcal{S}_2 = \{$ 4-tuples with two distinct values in pairs, i.e. with multiplicities $2, 2$ $\}$
3. $\mathcal{S}_3 = \{$ 4-tuples with two distinct values, multiplicities $3, 1$ $\}$
4. $\mathcal{S}_4 = \{$ 4-tuples with three distinct values, multiplicities $2, 1, 1$ $\}$
5. $\mathcal{S}_5 = \{$ 4-tuples with four distinct values, multiplicities $1, 1, 1, 1$ $\}$

Suppose that for a particular (s, t, v, w) one of the indices is different from all of the others. Without loss of generality, we may suppose this distinguished index is s. By the assumption it has the property that $s \neq t, v, w$. Then we would have by independence of the u's that $E(u_{is}u_{it}u_{iv}u_{iw}) = E(u_{is})E(u_{it}u_{iv}u_{iw}) = 0$. By this argument, for any (s, t, v, w) in $\mathcal{S}_3$, $\mathcal{S}_4$ or $\mathcal{S}_5$, we have $E(u_{is}u_{it}u_{iv}u_{iw}) = 0$, since for a 4-tuple in any of these sets there is always an index whose value has multiplicity one. Therefore we can rewrite (2.43) as

$$
\begin{aligned}
E\left\{(\mathbf{z}_{ij}^T\mathbf{u}_i)^2(\mathbf{z}_{ik}^T\mathbf{u}_i)^2\right\} &= \sum_{s=1}^{r} z_{ij}^{(s)2} z_{ik}^{(s)2} E(u_{is}^4) \\
&\quad + \sum_{(s,t,v,w)\in\mathcal{S}_2} z_{ij}^{(s)} z_{ij}^{(t)} z_{ik}^{(v)} z_{ik}^{(w)} E(u_{is}u_{it}u_{iv}u_{iw}) \\
&= \sum_{s=1}^{r} z_{ij}^{(s)2} z_{ik}^{(s)2} E(u_{is}^4) \\
&\quad + \sum_{1\leq a<b\leq r} E(u_{ia}^2 u_{ib}^2) \left\{ z_{ij}^{(a)2} z_{ik}^{(b)2} \right. \\
&\qquad \left. + 4 z_{ij}^{(a)} z_{ij}^{(b)} z_{ik}^{(a)} z_{ik}^{(b)} + z_{ij}^{(b)2} z_{ik}^{(a)2} \right\}, \qquad (2.44)
\end{aligned}
$$

where the second line follows by considering all $(s, t, v, w) \in \mathcal{S}_2$ such that $\{s, t, v, w\} = \{a, b\}$, $a < b$. Let us recall that for a normal random variable $\zeta \sim N(0, \sigma^2)$ the lower order moments are $E(\zeta^2) = \sigma^2$, $E(\zeta^3) = 0$ and $E(\zeta^4) = 3\sigma^4$. Using these we deduce that (2.44) can in fact be written as

$$
\begin{aligned}
E\left\{(\mathbf{z}_{ij}^T\mathbf{u}_i)^2(\mathbf{z}_{ik}^T\mathbf{u}_i)^2\right\} &= 3\sum_{s} z_{ij}^{(s)2} z_{ik}^{(s)2} G_{ss}^2 \\
&\quad + \sum_{a<b} \left\{ z_{ij}^{(a)2} z_{ik}^{(b)2} + z_{ij}^{(b)2} z_{ik}^{(a)2} + 4 z_{ij}^{(a)} z_{ij}^{(b)} z_{ik}^{(a)} z_{ik}^{(b)} \right\} G_{aa} G_{bb}, \qquad (2.45)
\end{aligned}
$$

Equation (2.45) can further be re-expressed as:

$$
\begin{aligned}
E\left\{(\mathbf{z}_{ij}^T\mathbf{u}_i)^2(\mathbf{z}_{ik}^T\mathbf{u}_i)^2\right\} &= 3\sum_{s} z_{ij}^{(s)2} z_{ik}^{(s)2} G_{ss}^2 \\
&\quad + \sum_{s\neq t} \left\{ z_{ij}^{(s)2} z_{ik}^{(t)2} + 2 z_{ij}^{(s)} z_{ij}^{(t)} z_{ik}^{(s)} z_{ik}^{(t)} \right\} G_{ss} G_{tt} \\
&= \sum_{s,t} \left\{ z_{ij}^{(s)2} z_{ik}^{(t)2} + 2 z_{ij}^{(s)} z_{ij}^{(t)} z_{ik}^{(s)} z_{ik}^{(t)} \right\} G_{ss} G_{tt} \\
&= (\mathbf{z}_{ij}^T G \mathbf{z}_{ij})(\mathbf{z}_{ik}^T G \mathbf{z}_{ik}) + 2(\mathbf{z}_{ij}^T G \mathbf{z}_{ik})^2 .
\end{aligned}
$$

Therefore $E(U_j'U_k') = \mathbf{z}_{ij}G\mathbf{z}_{ik}^T + (\lambda/4)\left\{(\mathbf{z}_{ij}^T G\mathbf{z}_{ij})(\mathbf{z}_{ik}^T G\mathbf{z}_{ik}) + 2(\mathbf{z}_{ij}^T G\mathbf{z}_{ik})^2\right\}$, where λ is the ratio of derivatives of h, $\lambda = \frac{h''(\eta_{ij})h''(\eta_{ik})}{h'(\eta_{ij})h'(\eta_{ik})}$. Since $E(U_j') = \frac{h''(\eta_{ij})}{2h'(\eta_{ij})}E\{(\mathbf{z}_{ij}^T u)^2\} = \frac{h''(\eta_{ij})}{2h'(\eta_{ij})}\mathbf{z}_{ij}^T G\mathbf{z}_{ij}$, we have using (2.41) that

$$\begin{aligned}\operatorname{cov}(U_j', U_k') &= \mathbf{z}_{ij}^T G\mathbf{z}_{ik} + (\lambda/4)\left\{(\mathbf{z}_{ij}^T G\mathbf{z}_{ij})(\mathbf{z}_{ik}^T G\mathbf{z}_{ik}) + 2(\mathbf{z}_{ij}^T G\mathbf{z}_{ik})^2\right\}\\ &\qquad - (\lambda/4)(\mathbf{z}_{ij}^T G\mathbf{z}_{ij})(\mathbf{z}_{ik}^T G\mathbf{z}_{ik})\\ &= \mathbf{z}_{ij}^T G\mathbf{z}_{ik} + (\lambda/2)(\mathbf{z}_{ij}^T G\mathbf{z}_{ik})^2\\ &= \mathbf{z}_{ij}^T G\mathbf{z}_{ik} + (1/2)\frac{h''(\eta_{ij})h''(\eta_{ik})}{h'(\eta_{ij})h'(\eta_{ik})}(\mathbf{z}_{ij}^T G\mathbf{z}_{ik})^2\,.\end{aligned}$$

Note that the above corresponds with the expression we found for $\operatorname{var}(t_{ij})$.

We now write the variance matrix for the whole block, $\operatorname{var}(\mathbf{t}_i)$, in matrix form. Let us denote by Z_i the matrix whose columns are the $\mathbf{z}_{ij}$. Then, since G is diagonal, the matrix $Z_i^T G Z_i$ has (j,k)th entry $\mathbf{z}_{ij}^T G\mathbf{z}_{ik}$, so that we may write

$$\operatorname{var}(\mathbf{t}_i) = \operatorname{diag}\left\{\frac{1}{h'(\eta_{ij})} : 1 \le j \le m_i\right\} + Z_i^T G Z_i + (1/2)\frac{h''(\boldsymbol{\eta}_i)}{h'(\boldsymbol{\eta}_i)}\left(\frac{h''(\boldsymbol{\eta}_i)}{h'(\boldsymbol{\eta}_i)}\right)^T * (Z_i^T G Z_i)^{*2}\,,$$

where $*$ denotes componentwise (Hadamard) multiplication of matrices, and the vector $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{im})^T$. The definition of the (originally scalar) derivative functions, h' and h'', is here extended to permit vector arguments by acting componentwise, e.g. $h'(\boldsymbol{\eta}_i) = (h'(\eta_{i1}), \ldots, h'(\eta_{im}))^T$. For vectors $\mathbf{a}, \mathbf{b}$ of the same length we define $\frac{\mathbf{a}}{\mathbf{b}}$ to be the result of componentwise division.

2.10 Appendix: The integration method of Gotwalt et al. (2009)

2.10.1 Adaptation to the logistic random intercept model

Gotwalt et al. (2009) proposed a numerical method for evaluating the log-determinant objective function of Chaloner and Larntz (1989),

$$I_0(\xi) = \int_{\mathbb{R}^p} \log|M_{\boldsymbol{\beta}}(\xi, \boldsymbol{\beta})|\, d\mathcal{P}(\boldsymbol{\beta})\,,$$

under the logistic model with parameters $\boldsymbol{\beta}$ when the prior, $\mathcal{P}$, on $\boldsymbol{\beta}$ is multivariate normal. Here the method is adapted to evaluate the objective function $I_{1/p}(\xi)$ of Firth and Hinde (1997), under the logistic random intercept model of Section 2.1. As the approximation lends itself to the use of normal priors, we shall adopt the normal as a default prior on the fixed effects parameters. However, the random effects variance must be positive and so for this we use a log-normal prior. We shall assume a priori that the parameters are independent.

The derivation of Gotwalt et al. (2009) applies equally well if a function other than $\log|M|$ is used in the integrand, provided the distribution on the integration variables remains the same. Therefore the approximation can be restated in the following way: if $\mathcal{D}$ is a multivariate normal distribution of dimension d, with mean $\bar{\mathbf{t}}$ and variance Σ, and ψ is a general function of $\mathbf{t}$, a placeholder variable, then

$$\int \psi(\mathbf{t})\, d\mathcal{D}(\mathbf{t}) \approx w_{R_0}\psi(\bar{\mathbf{t}}) + \sum_{i=1}^{n_R}\sum_{j=1}^{n_Q}\sum_{k=1}^{n_S} \frac{w_{R_i}w_{S_k}}{n_Q}\, \psi(\bar{\mathbf{t}} + L\sqrt{\tau_i}Q_{ij}\mathbf{v}_k)\,, \tag{2.46}$$

where $\{\tau_i, w_{R_i}\}_{i=0}^{n_R}$ are the generalised Gauss-Laguerre abscissae and weights (Cassity, 1965), $\{\mathbf{v}_k, w_{S_k}\}_{k=0}^{n_S}$ are the abscissae and weights from the Mysovskikh extended simplex rule, and the $\{Q_{ij} : 1 \leq i \leq n_R,\ 1 \leq j \leq n_Q\}$ are randomly generated orthogonal matrices which rotate the Mysovshikh simplex. The matrix L is the lower Cholesky root of Σ, so that $\Sigma = LL^T$. The various abscissae and weights will be defined in the next subsection, however for the moment note that if $\mathbf{t}^T = (\mathbf{t}_1^T, t_2) = (\boldsymbol{\beta}^T, \log\sigma^2)$, with β and σ^2 the logistic random intercept model parameters, and

$$\begin{aligned}\psi(\mathbf{t}) &= |M_{\boldsymbol{\beta}}(\xi,\boldsymbol{\theta})|^{1/p}\,,\\ &= \left|M_{\boldsymbol{\beta}}\left(\xi,(\mathbf{t}_1^T, e^{t_2})^T\right)\right|^{1/p}\,,\end{aligned} \tag{2.47}$$

then the LHS of (2.46) becomes

$$I_{1/p}(\xi) = \int |M(\xi,\boldsymbol{\theta})|^{1/p}\, d\mathcal{P}(\boldsymbol{\theta})\,, \tag{2.48}$$

where $\mathcal{P}$ is the normal-lognormal prior on $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2)^T$ which is induced by the distribution $\mathcal{D}$ on $\mathbf{t}$. This is seen from the fact that if V and W are random variables related by $V = \Psi(W)$, with Ψ an arbitrary fixed function, then the expectations $E(V)$ and $E\{\Psi(W)\}$ are identical. Therefore, to evaluate (2.48) approximately, substitute the function ψ from (2.47) into (2.46).

2.10.2 Abscissae and weights

The radial abscissae, τ_i, $i = 1, \ldots, n_R$, are given by $2a_i$, where the a_i are the roots of the generalised Laguerre polynomial of degree n_R with parameter $(p+1)/2$. The parameter differs slightly from that in the result of Gotwalt et al. (2009), since in their paper the dimension of the integral was p whilst in our case the dimension of (2.48) is $p+1$. The roots of the generalised Laguerre polynomial can be found using the algorithm of Press, Teuklosky, Vetterling and Flannery (1992, pp.147-151). The corresponding weights are

$$w_i = \frac{\Gamma(n_R+1)\Gamma(n_R+(p+1)/2)}{(n_R+(p+1)/2)\Gamma((p+1)/2)\{L_{n_R-2}^{(p+1)/2-1}(a_i)\}^2}\,,$$

where L_n^s denotes the generalised Laguerre polynomial of order n with parameter s (Cassity, 1965), and Γ denotes the gamma function. The construction of the Mysovkikh extended simplex requires several steps. First of all one creates a simplex of $p+2$ vertices $\mathbf{v}_i$, $i = 0, \ldots, p+1$, on the unit sphere in $\mathbb{R}^{p+1}$. This simplex is defined by $\mathbf{v}_i = (v_{i1}, \ldots, v_{i(p+1)})^T$ and

$$v_{ij} = \begin{cases} -\sqrt{\frac{p+2}{(p+1)(p-j+3)(p-j+2)}}\,, & j < i \\ \sqrt{\frac{(p+2)(p-i+2)}{(p+1)(p-i+3)}}\,, & j = i \\ 0\,, & j > i\,. \end{cases}$$

Then the midpoints of these vertices are added to the simplex. The construction is then completed by adding the negatives of all vertices and midpoints. The weights corresponding to the vertices and their negatives are all equal to $(p+1)(6-p)/\{2(p+2)^2(p+3)\}$, and the weights corresponding to midpoints and their negatives are all equal to $2p^2/\{(p+1)(p+2)^2(p+3)\}$.

We obtain the matrices Q_{ij} by randomly generating matrices whose entries have independent standard normal distributions, and then taking the QR factorisation, as mentioned by, for example, Stewart (1980). We implemented this using the built in functions in the programming language `R`.

Chapter 3

Comparison of approximations

In this chapter, we compare the ability of the different approximations in Chapter 2 to produce efficient designs for the logistic random intercept model. To perform the comparison, we use a few different benchmarks. In Section 3.1, we examine the sensitivity of the MQL and PQL approximations to the use of different allocations of the same treatments among blocks. The relative D-efficiencies of different allocations are computed under each of the approximations, and compared with the efficiencies computed numerically using complete enumeration. Section 3.2 develops a methodology, which we call maximum likelihood by numerical interpolation (MLNI), for computing the optimal design under ML in the case where there are two points per block. This technique is also extended to yield Bayesian designs without having to resort to the less accurate analytical approximations to the information matrix. In Section 3.3 we propose a further analytical approximation (AMQL), which is superior to those in Chapter 2. Despite being computationally very cheap, AMQL performs comparably with MLNI in the examples of Section 3.4.

3.1 Detection of optimal allocation

In this section, we take two different allocations of a four-point GLM design (corresponding to $\sigma^2 = 0$) and evaluate the relative efficiency of the two allocations using various approximations. The initial design is a four-point exact D-optimal design for the logistic regression model with two factors x_1 and x_2 and linear predictor

$$\eta(\mathbf{x};\boldsymbol{\beta}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2\,, \tag{3.1}$$

with $(\beta_0, \beta_1, \beta_2)^T = (0, 5, 10)^T$. The locally optimal design points, $\{A, B, C, D\} \subseteq [-1, 1]^2$, for this problem are shown in Figure 3.1. We consider allocating these four points to an approximate block design, in the sense of Section 2.1.3, in two different ways:

1. Allocation 1: *Block 1,* $\zeta_1 = \{A, B\}$. *Block 2,* $\zeta_2 = \{C, D\}$. We expect this to be a poor allocation because the level of x_1 is constant within each block. Hence we anticipate some confounding between the effect of x_1 and the block effects.

2. Allocation 2: *Block 1,* $\zeta_1 = \{A, C\}$. *Block 2,* $\zeta_2 = \{B, D\}$. We expect that this offers some

improvement over allocation 1 above, since within each block there are two levels of both x_1 and x_2.

In both cases, the blocks are to be equally weighted.

The resulting approximate block designs are to be used to estimate the logistic random intercept model with conditional linear predictor

$$\nu(\mathbf{x}|u) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u\,,$$

where the true values of the fixed effects parameters are $(\beta_0, \beta_1, \beta_2)^T = (0, 5, 10)^T$ in accordance with the GLM above. The relative efficiency of Allocation 1, using Allocation 2 as a reference has been computed using the analytical approximations of Chapter 2 (MQL, PQL, and MQL2), complete enumeration (which we refer to simply as ML), and the improved analytical approximation, adjusted MQL (AMQL). We defer the details of the AMQL approximation to Section 3.3, but include it in the considerations here for the sake of completeness. The efficiencies under each of these methods are displayed as functions of σ^2 in Figure 3.2.

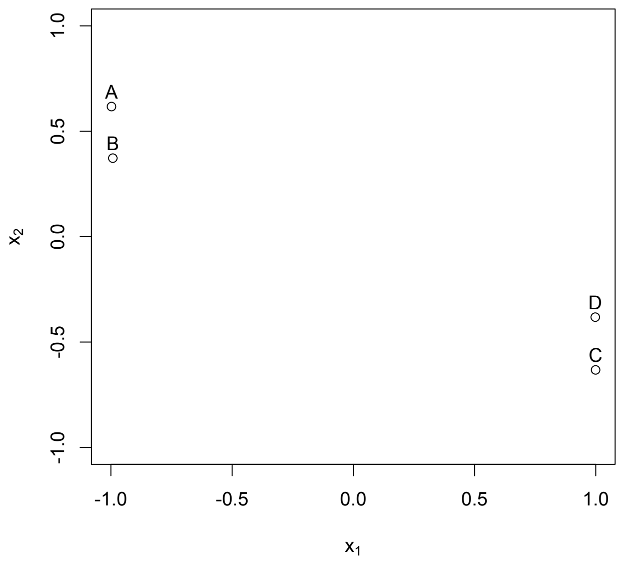

Figure 3.1: D-optimal design on 4 points for the model (3.1).

We see that the correct pattern, given by the ML curve, is for the relative efficiency of Allocation 1 to decrease monotonically as σ^2 increases. Therefore the choice of allocation becomes more important as the degree of heterogeneity among the blocks increases. This pattern is also followed by the MQL and adjusted MQL (AMQL) efficiencies, although these approximations tend to underestimate the efficiency of Allocation 1, thereby exaggerating the importance of selecting Allocation 2. The information matrix evaluations necessary to compute the relative ML efficiencies were performed using complete enumeration, as in (2.9). For values of σ^2 up to around 1, the different methods produce similar efficiency curves.

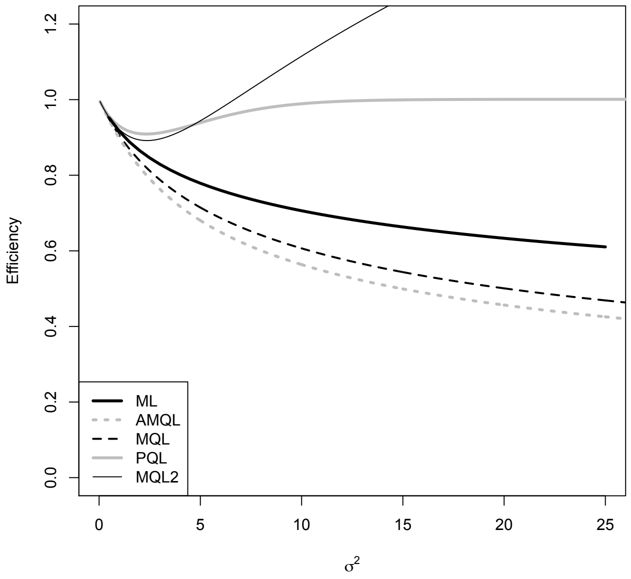

Figure 3.2: Relative D-efficiency of Allocation 1, under various approximations

For $\sigma^2 \geq 1$ the approximations begin to diverge noticeably. Under PQL the efficiency decreases to a minimum of roughly 90% at $\sigma^2 \approx 3$. For $\sigma^2 \geq 3$, the efficiency increases back to 1 and remains at that level. In other words, according to PQL, for moderately large σ^2 there is no difference between the two allocations. In Section 3.1.1, we will show that PQL is always insensitive to allocation for large σ^2. This appears to be a serious defect for an approximate method of computing block designs. Moreover, as we shall see in Sections 3.4 and 4.4.5, for moderately large σ^2 the treatments in the optimal PQL design tend to be worse even than those from a GLM design, in other words a design completely ignoring the presence of a block effect. Thus, it seems PQL is a worse approximation than MQL for the purposes of design construction.

MQL2 is the only approximation which selects the wrong allocation, doing so for σ^2 greater than around 7. Furthermore, the efficiency is very much greater than 1 for large σ^2. This suggests that MQL2 is the worst approximation, which is perhaps initially surprising given its second order nature. However recall that the conditional mean, $h(\nu_{ij})$, lies between 0 and 1. As outlined in Section 2.3.6, MQL2 attempts to approximate this quantity with an expression which contains second order terms in the random effects. When σ^2 is large, the terms involving random effects will often dominate, causing the approximation to lie outside the bounds. Moreover, the second order terms in MQL2 will tend to be larger than the first order terms present in MQL and PQL. As a result MQL2 will perform worse when σ^2 is large, in virtue of the approximation to $h(\nu_{ij})$ lying further from [0,1].

3.1.1 Analytical results

In this section, we study the behaviour of the MQL and PQL information matrices of a fixed design as $\sigma^2 \to \infty$. This leads to an analytical proof of the property identified in Section 3.1, that PQL is insensitive to the allocation of treatments to blocks for large σ^2. Of course, for very large σ^2 the model is degenerate: as stated in Section 2.1.1, in this case within most blocks the responses will be equal. Thus it will not matter much in practice which allocation is chosen as either way the inference will be very poor. However, as in Section 3.1, low sensitivity to the allocation used may occur for intermediate values of σ^2, for which the model is not yet degenerate.

Recall from Section 2.3 that the MQL and PQL information matrices for a design ξ can be written in the form

$$M(\xi; \boldsymbol{\theta}) = \sum_{i=1}^{b} w_i F_i^T V_i^{-1} F_i \,, \tag{3.2}$$

where F_i is the model matrix of the ith block, ζ_i, in the design. Also, w_i is the corresponding weight, and V_i is the variance-covariance matrix of the working variate in block i which depends on the approximation method. We shall invert V_i analytically using the following matrix formula (see Fedorov and Hackl, 1997, p. 107).

Lemma 3.1 (Inversion and determinant formula)**.** *Let A be an invertible $a \times a$ matrix, and B an $a \times b$ matrix. Then we have the following expressions for the inverse and determinant of $A + BB^T$,*

$$\left|A + BB^T\right| = |A| \left|I + B^T A^{-1} B\right| \tag{3.3}$$

$$(A + BB^T)^{-1} = A^{-1} - A^{-1}B(I + B^T A^{-1} B)^{-1} B^T A^{-1} \,. \tag{3.4}$$

The next result is useful for showing that PQL is insensitive to the allocation chosen for large values of the random effects variance, σ^2. Note that a corollary of this is the intuitive fact that if the observations within a block are independent, then the allocation of treatments to blocks does not matter.

Lemma 3.2. *If the V_i, $i = 1, \ldots, b$ in (3.2) are diagonal, then M is insensitive to the choice of allocation, provided each point is moved to a block with the same weight as before.*

Proof of Lemma 3.2. As V_i is diagonal we may write it as

$$V_i = \begin{pmatrix} v_{i1} & & & \\ & v_{i2} & & \\ & & \ddots & \\ & & & v_{im} \end{pmatrix} .$$

Note that, by the matrix algebra of outer products

$$\begin{aligned} F_i^T V_i^{-1} F_i &= \begin{pmatrix} \mathbf{f}(\mathbf{x}_{i1}) & \mathbf{f}(\mathbf{x}_{i2}) & \ldots & \mathbf{f}(\mathbf{x}_{im}) \end{pmatrix} \begin{pmatrix} v_{i1}^{-1} & & & \\ & v_{i2}^{-1} & & \\ & & \ddots & \\ & & & v_{im}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{f}^T(\mathbf{x}_{i1}) \\ \mathbf{f}^T(\mathbf{x}_{i2}) \\ \vdots \\ \mathbf{f}^T(\mathbf{x}_{im}) \end{pmatrix} \\ &= \sum_{j=1}^{m} v_{ij}^{-1} \mathbf{f}(\mathbf{x}_{ij}) \mathbf{f}^T(\mathbf{x}_{ij}) \,. \end{aligned}$$

This can be verified by writing out all the necessary indexed sums, if desired.

Note therefore that

$$M(\xi;\theta) = \sum_{i=1}^{b}\sum_{j=1}^{m} w_i v_{ij}^{-1}\mathbf{f}(\mathbf{x}_{ij})\mathbf{f}^T(\mathbf{x}_{ij})\,.$$

This remains unchanged if the indices i and j are permuted between the points $\mathbf{x}_{ij}$, provided that the weight associated with a given point does not change. Thus, given the hypothesis, M is insensitive to the chosen allocation, subject to the stated caveat.

□

Large σ^2 properties of MQL

For MQL,

$$W_i = \text{diag}\left\{\frac{1}{\mu_{ij}^{(0)}(1-\mu_{ij}^{(0)})} \;:1 \le j \le m\right\}\,,$$

where diag$\{d_j\ : 1 \le j \le m\}$ denotes the $m \times m$ diagonal matrix with diagonal entries d_j, and $\mu_{ij}^{(0)} = \mu(\mathbf{x}_{ij}|0)$, in other words $\mu_{ij}^{(0)}$ is the conditional mean of the response assuming that the random effect is equal to 0. Then for MQL, the variance matrix V_i is given by

$$V_i = W_i + \sigma^2\mathbf{1}_m\mathbf{1}_m^T\,, \tag{3.5}$$

where $\mathbf{1}_m$ is an $m \times 1$ vector of 1s. Applying formula (3.4) to equation (3.5), by setting $A = W_i$ and $B = \sigma\mathbf{1}_m$, yields

$$\begin{aligned} V_i^{-1} &= W_i^{-1} - W_i^{-1}(\sigma\mathbf{1}_m)\left\{I + (\sigma\mathbf{1}_m)^T W_i^{-1}(\sigma\mathbf{1}_m)\right\}^{-1}(\sigma\mathbf{1}_m^T)W_i^{-1} \\ &= W_i^{-1} - \sigma^2 W_i^{-1}\mathbf{1}_m\left\{1 + \sigma^2\text{tr}(W_i^{-1})\right\}^{-1}\mathbf{1}_m^T W_i^{-1} \\ &= W_i^{-1} - \frac{\sigma^2}{1+\sigma^2\text{tr}(W_i^{-1})}W_i^{-1}\mathbf{1}_m\mathbf{1}_m^T W_i^{-1T} \\ &= W_i^{-1} - \frac{1}{\sigma^{-2}+\text{tr}(W_i^{-1})}W_i^{-1}\mathbf{1}_m(W_i^{-1}\mathbf{1}_m)^T\,. \end{aligned} \tag{3.6}$$

Note that $W_i^{-1}\mathbf{1}_m$ is the vector containing the diagonal elements of W_i^{-1}, in other words $W_i^{-1}\mathbf{1}_m = (\,\mu_{ij}^{(0)}(1-\mu_{ij}^{(0)})\ :\ 1 \le j \le m\,)^T$. Considering the RHS of equation (3.6) we see that, as $\sigma^2 \to \infty$,

$$V_i^{-1} \to W_i^{-1} - \frac{1}{\text{tr}(W_i^{-1})}W_i^{-1}\mathbf{1}_m(W_i^{-1}\mathbf{1}_m)^T\,.$$

Defining $V_{i,\infty}^{-1} = \lim_{\sigma^2\to\infty} V_i^{-1}$, to evaluate the limiting efficiency two designs ξ_1 and ξ_2 one needs only evaluate

$$\lim_{\sigma^2\to\infty} M(\xi,\theta) = \sum_{i=1}^{b} w_i F_i^T V_{i,\infty}^{-1} F_i\,,$$

for each of the designs under consideration. The matrix $V_{i,\infty}^{-1}$ is not diagonal, so the allocation potentially remains important in the limit.

Large σ^2 properties of PQL

For PQL

$$W_i = \text{diag}\left\{2 + 2e^{\sigma^2/2}\cosh(\mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta})\ :\ 1 \le j \le m\right\}\,,$$

and note that

$$V_i = W_i + \sigma^2 \mathbf{1}_m \mathbf{1}_m^T\,.$$

It is helpful to consider a rescaled version of V_i, namely

$$\tilde{V}_i = e^{-\sigma^2/2} V_i\,. \tag{3.7}$$

Correspondingly we consider a rescaled version of W_i, defined by

$$\begin{aligned}\tilde{W}_i &= e^{-\sigma^2/2} W_i \\ &= \operatorname{diag}\left\{2e^{-\sigma^2/2} + 2\cosh(\mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta}) \,:\, 1 \le j \le m\right\}\,,\end{aligned}$$

so that $\tilde{V}_i = \tilde{W}_i + \sigma^2 e^{-\sigma^2/2}\mathbf{1}_m\mathbf{1}_m^T$. Clearly, as $\sigma^2 \to \infty$,

$$\tilde{W}_i \to \tilde{W}_{i,\infty} = \operatorname{diag}\left\{2\cosh(\mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta}) \,:\, 1 \le j \le m\right\}.$$

Using (3.7) we can rewrite the PQL information matrix as

$$\begin{aligned}M(\xi,\theta) &= e^{\sigma^2/2}\sum_{i=1}^{b} w_i F_i^T \tilde{V}_i^{-1} F_i \\ &= e^{\sigma^2/2}\tilde{M}(\xi,\boldsymbol{\theta})\,,\end{aligned}$$

where we refer to $\tilde{M}$ as the *renormalised (PQL) information matrix.* Applying the inversion formula (3.4) to $\tilde{V}_i$ gives

$$\begin{aligned}\tilde{V}_i^{-1} &= \tilde{W}_i^{-1} - \frac{\sigma^2 e^{-\sigma^2/2}}{1+\sigma^2 e^{-\sigma^2/2}\operatorname{tr}(\tilde{W}_i^{-1})}\tilde{W}_i^{-1}\mathbf{1}_m(\tilde{W}_i^{-1}\mathbf{1}_m)^T \\ &= \tilde{W}_i^{-1} - \frac{1}{\sigma^{-2}e^{\sigma^2/2}+\operatorname{tr}(\tilde{W}_i^{-1})}\tilde{W}_i^{-1}\mathbf{1}_m(\tilde{W}_i^{-1}\mathbf{1}_m)^T\,.\end{aligned} \tag{3.8}$$

As $\sigma^2 \to \infty$, so too $\sigma^{-2}e^{\sigma^2/2} \to \infty$. Moreover, since $\tilde{W}_i \to \tilde{W}_{i,\infty}$ so too $\tilde{W}_i^{-1} \to \tilde{W}_{i,\infty}^{-1}$ and also $\operatorname{tr}(\tilde{W}_i^{-1}) \to \operatorname{tr}(\tilde{W}_{i,\infty}^{-1})$ which is a fixed real number. Therefore the second term in (3.8) tends to

$$\lim_{\sigma^2\to\infty}\left(\frac{1}{\sigma^{-2}e^{\sigma^2/2}+\operatorname{tr}(\tilde{W}_i^{-1})}\right)\tilde{W}_{i,\infty}^{-1}\mathbf{1}_m(\tilde{W}_{i,\infty}^{-1}\mathbf{1}_m)^T = 0_{m\times m}\,,$$

and $\tilde{V}_i^{-1} \to W_{i,\infty}^{-1}$. Thus, as $\sigma^2 \to \infty$, $\tilde{V}_i^{-1}$ tends to a diagonal matrix.

When calculating efficiencies, the renormalised information matrix may be used, since the normalisation factors cancel in the efficiency equation as follows:

$$\begin{aligned}\operatorname{eff}(\xi_1;\xi_2,\boldsymbol{\theta}) &= \left\{\frac{|M(\xi_1,\boldsymbol{\theta})|}{|M(\xi_2,\boldsymbol{\theta})|}\right\}^{1/p} \\ &= \left\{\frac{e^{\sigma^2/2}|\tilde{M}(\xi_1;\boldsymbol{\theta})|}{e^{\sigma^2/2}|\tilde{M}(\xi_2;\boldsymbol{\theta})|}\right\}^{1/p} \\ &= \left\{\frac{|\tilde{M}(\xi_1,\boldsymbol{\theta})|}{|\tilde{M}(\xi_2,\boldsymbol{\theta})|}\right\}^{1/p}\,.\end{aligned}$$

Since for large σ^2 the renormalised information matrix is independent of the allocation, so the relative efficiency of two allocations of the same design must be 1. In other words, for large σ^2, PQL does not distinguish between allocations.

In Section 3.4, we observe that this allocation property is not the only deficiency of PQL. It can also produce worse design points than MQL (and even than we would obtain simply by ignoring the presence of random effects). This is somewhat counterintuitive, since as an estimation procedure PQL has been found to yield better results than MQL. The answer may lie in the gross approximations we make in the derivations in Sections 2.3.5 and 2.8 to guarantee an analytically tractable expression for $\mathrm{var}(\mathbf{t}_i)$. These approximations are of higher fidelity when the random effects variability is small, which we know since PQL and MQL expressions are identical to those in the GLM case when $\sigma^2 = 0$. For larger σ^2 there is no reason that these steps should be accurate. To resolve whether the deficiencies in our PQL are down to these steps, one could consider a Monte Carlo PQL expression along the lines of Tekle et al. (2008). However, doing so would reduce the computational advantages over complete enumeration.

3.2 ML design methodology

In this section we have two objectives. Firstly, we present a faster method of evaluating the information matrix using complete enumeration (2.9). Secondly we discuss a methodology, referred to as maximum likelihood by numerical interpolation (MLNI), which enables us to obtain D-optimal designs under ML in the case where there are two units per block. The key idea is to use numerical integration to precompute a lookup table for the *weight matrix*, W (Section 3.2.1). This table will contain practically all numerical integrals which are possibly relevant to the computation of M for a particular value of σ^2.

The weight matrix, W, will turn out to depend only on the values of the fixed parts of the linear predictor at the two points in the block. Therefore the same lookup table can be used for different predictor structures and, perhaps most importantly, different values of the parameters $\boldsymbol{\beta}$. This property facilitates the construction of Bayesian designs when there is uncertainty only in the $\boldsymbol{\beta}$ parameters, as will be considered in Section 3.4.

3.2.1 Alternate expression for information matrix

Recall that the information matrix for a block ζ can be written as a sum over all possible outcomes in the block of terms involving the likelihood $p(\mathbf{y}|\zeta, \boldsymbol{\theta})$ and its derivatives with respect to $\boldsymbol{\beta}$. Previously we computed $p(\mathbf{y}|\zeta, \boldsymbol{\theta})$ using quadrature to evaluate the integral, and numerical differentiation to approximate the Hessian. Here we show how to avoid the use of numerical differentiation.

The first step is to make a change of differentiation variable from $\boldsymbol{\beta}$ to $\boldsymbol{\eta}$ using the chain rule. The second step is to use differentiation under the integral sign to express the derivative with respect to $\boldsymbol{\eta}$ of the likelihood in terms of an integral. A full description of differentiation under the integral, together with a list of conditions on the integrand to enable its use, are given in Section 3.9.

Let $\zeta = (\mathbf{x}_1, \dots \mathbf{x}_m) \in \mathcal{X}^m$ be an arbitrary block. From (2.9), we have that

$$M_{\boldsymbol{\beta}}(\zeta, \boldsymbol{\theta}) = \sum_{\mathbf{y} \in \{0,1\}^m} \frac{-\partial^2 \log p(\mathbf{y}|\boldsymbol{\theta}, \zeta)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}} p(\mathbf{y}|\boldsymbol{\theta}, \zeta) \,.$$

Using standard ML theory, under regularity this can be rewritten as

$$M_{\boldsymbol{\beta}}(\zeta, \boldsymbol{\theta}) = \sum_{\mathbf{y} \in \{0,1\}^m} p(\mathbf{y}|\boldsymbol{\theta}, \zeta) \left(\frac{\partial \log p}{\partial \boldsymbol{\beta}} \right) \left(\frac{\partial \log p}{\partial \boldsymbol{\beta}} \right)^T ,$$

which can be further rewritten, using the chain rule on $\log p$, as

$$M_{\boldsymbol{\beta}}(\zeta, \boldsymbol{\theta}) = \sum_{\mathbf{y} \in \{0,1\}^m} \frac{1}{p(\mathbf{y}|\boldsymbol{\theta}, \zeta)} \frac{\partial p}{\partial \boldsymbol{\beta}} \frac{\partial p}{\partial \boldsymbol{\beta}}^T \,. \tag{3.9}$$

Let $\boldsymbol{\eta} = (\eta_1, \dots, \eta_m)^T$ be the vector of linear predictors, $\eta_j = \mathbf{f}^T(\mathbf{x}_j)\boldsymbol{\beta}$, $j = 1, \dots, m$, and let $\mathbf{z}_j = \mathbf{z}(\mathbf{x}_j)$. Moreover write $h = g^{-1}$ for the inverse link function. Recall from (2.8) that the probability of a particular outcome for the block, $\mathbf{y}$, is

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\theta}, \zeta) &= \int_{\mathbb{R}^r} \prod_{j=1}^m \mu(\mathbf{x}_j|\mathbf{u})^{y_j} \{1 - \mu(\mathbf{x}_j|\mathbf{u})\}^{(1-y_j)} f_{\mathbf{u}}(\mathbf{u}) \, d\mathbf{u} \\ &= \int_{\mathbb{R}^r} \prod_{j=1}^m \Big[y_j h(\eta_j + \mathbf{z}_j^T \mathbf{u}) + (1 - y_j)\{1 - h(\eta_j + \mathbf{z}_j^T \mathbf{u})\} \Big] f_{\mathbf{u}}(\mathbf{u}) \, d\mathbf{u} \,. \end{aligned} \tag{3.10}$$

To see that the second expression in (3.10) is correct, one simply needs to consider the possible cases, $y_j \in \{0, 1\}$, $1 \le j \le m$, and verify that in each eventuality we obtain the same integrand as in the first line. The purpose of expressing the integrand using linear combinations rather than exponentiation is that, in R at least, the former is computationally faster. Moreover, it is more straightforward to analytically differentiate linear combinations.

Note that (3.10) depends on the parameters $\boldsymbol{\beta}$ only through the linear predictors $\boldsymbol{\eta}$. To stress this, we write

$$p(\mathbf{y}|\boldsymbol{\theta}, \zeta) = p_{\mathbf{y}}(\boldsymbol{\eta}, \zeta, G) \,,$$

where $G = \text{var}(\mathbf{u})$. By the chain rule, we have that

$$\frac{\partial p_{\mathbf{y}}}{\partial \boldsymbol{\beta}} = F^T \frac{\partial p_{\mathbf{y}}}{\partial \boldsymbol{\eta}} \,,$$

where F is the $m \times p$ model matrix for the block ζ, which is given by stacking the $\mathbf{f}^T(\mathbf{x}_j)$, $j = 1, \dots, m$. Thus in fact

$$\begin{aligned} M_{\boldsymbol{\beta}}(\zeta, \boldsymbol{\theta}) &= F^T \left\{ \sum_{\mathbf{y} \in \{0,1\}^m} \frac{1}{p_{\mathbf{y}}} \left(\frac{\partial p_{\mathbf{y}}}{\partial \boldsymbol{\eta}} \right) \left(\frac{\partial p_{\mathbf{y}}}{\partial \boldsymbol{\eta}} \right)^T \right\} F \\ &= F^T W F \,, \end{aligned} \tag{3.11}$$

where the $m \times m$ matrix $W = W(\boldsymbol{\eta}, \zeta, G)$ is defined by

$$W(\boldsymbol{\eta}, \zeta, G) = \sum_{\mathbf{y} \in \{0,1\}^m} \frac{1}{p_{\mathbf{y}}} \left(\frac{\partial p_{\mathbf{y}}}{\partial \boldsymbol{\eta}} \right) \left(\frac{\partial p_{\mathbf{y}}}{\partial \boldsymbol{\eta}} \right)^T . \tag{3.12}$$

We call W the *weight matrix.* In the case of a random intercept model, $p_{\mathbf{y}}$ and W depend upon ζ only through $\boldsymbol{\eta}$ and we can write $p_{\mathbf{y}} = p_{\mathbf{y}}(\boldsymbol{\eta}, \sigma^2)$ and $W = W(\boldsymbol{\eta}, \sigma^2)$. We will also sometimes notationally supress dependence on σ^2, since in the locally optimal design problem σ^2 is held fixed throughout.

Note that a numerical issue may potentially arise when evaluating (3.12) if p_y is too small. In practice this has not occurred in our work.

We now give expressions for the partial derivatives $\partial p_{\mathbf{y}} / \partial \boldsymbol{\eta}$. Differentiating under the integral sign in (3.10) we obtain for $1 \leq k \leq m$,

$$\begin{aligned} \frac{\partial p_{\mathbf{y}}}{\partial \eta_k} =& (2y_k - 1) \\ & \cdot \int_{\mathbb{R}^r} h'(\eta_k + \mathbf{z}_k^T \mathbf{u}) \prod_{j \neq k} \Big[y_j h(\eta_j + \mathbf{z}_j^T \mathbf{u}) \\ & \qquad + (1 - y_j)\{1 - h(\eta_j + \mathbf{z}_j^T \mathbf{u})\} \Big] f_{\mathbf{u}}(\mathbf{u}) \, d\mathbf{u} . \end{aligned} \tag{3.13}$$

We use quadrature, via the `R` function `integrate`, to evaluate (3.10) and (3.13) directly, and this enables us to evaluate the information matrix using (3.11) without resorting to numerical differentation (i.e. finite difference methods).

3.2.2 Maximum likelihood by numerical interpolation

We now outline the computational strategy for obtaining D-optimal ML designs for the logistic random intercept model. In our examples we restrict our attention to the case where there are $m = 2$ units per block. Further technical details on the implementation in this case are given in Section 3.2.3.

The idea is simple: given σ^2, let us suppose that we could construct tables of the values of the functions $p_{\mathbf{y}}(\boldsymbol{\eta})$, $\partial p_{\mathbf{y}} / \partial \eta_k (\boldsymbol{\eta})$, $k = 1, \ldots, m$, evaluated over a fine grid of values of $\boldsymbol{\eta}$ in $\mathbb{R}^m$. For our examples it was adequate to tabulate over the bounded region $[-20, 20]^m$, because the parameter space and the design space were both bounded. Once the function values have been obtained on the grid, subsequent evaluations of the information matrix can be performed almost instantaneously by 'looking up' the values of $p_{\mathbf{y}}$ and $\partial p_{\mathbf{y}} / \partial \eta_k$ at the value of $\boldsymbol{\eta}$ corresponding to the particular design of interest.

Of course, we will want to evaluate $p_{\mathbf{y}}$ and $\partial p_{\mathbf{y}} / \partial \eta_k$ at values of $\boldsymbol{\eta}$ other than those contained in the grid, but we can approximate the values of the functions at such non-grid points by a suitable multivariate interpolation method.

Because $p_{\mathbf{y}}$ and $\partial p_{\mathbf{y}} / \partial \eta_k$ depend on $\boldsymbol{\beta}$ only through $\boldsymbol{\eta}$, we can use the same interpolation tables no matter what the value of $\boldsymbol{\beta}$. We can also use the same tables in problems with different numbers of factors, and which include different functions of the factors in the regression (in other words $\mathbf{f}$ is also allowed to vary). The only restrictions are that if σ^2 or the block size are changed then the tables must be recalculated. This versatility will be useful in obtaining designs robust to model parameters (and structure), such as when computing Bayesian optimal designs in Section 3.4.2.

The difficulty in constructing interpolation tables of the type described above depends on the block size, m, in two ways. Firstly, for a grid with fixed step length, the number of points in the grid increases exponentially with the block size (the curse of dimensionality). Secondly, $p_{\mathbf{y}}$ and $\partial p_{\mathbf{y}}/\partial\eta_k$ must be tabulated for each potential outcome $\mathbf{y}$. There are 2^m such outcomes, and so as m increases the number of functions to be tabulated also increases exponentially. As a result, precomputation of interpolation tables is likely to be a feasible strategy only when m is small, say $m = 2, 3$.

We note a technical feature which applies for all m, which is that we do not need to tabulate $\partial p_{\mathbf{y}}/\partial\eta_k$ for all $\mathbf{y} \in \{0,1\}^m$ and all $1 \le k \le m$. Instead it will suffice to precompute only the first partial derivative, $\partial p_{\mathbf{y}}/\partial\eta_1$ for $\mathbf{y} \in \{0,1\}^m$. The reason for this is as follows. Let $\mathbf{y}^{(k)} = (y_k, y_2, \ldots, y_{k-1}, y_1, y_{k+1}, \ldots, y_m)^T$ denote the vector obtained by exchanging the first and kth components of $\mathbf{y}$, and similarly for $\boldsymbol{\eta}^{(k)}$. Then

$$\frac{\partial p_{\mathbf{y}}}{\partial\eta_k}(\boldsymbol{\eta}) = \frac{\partial p_{\mathbf{y}^{(k)}}}{\partial\eta_1}(\boldsymbol{\eta}^{(k)}), \tag{3.14}$$

and the second term can be evaluated using the tables for $\partial p_{\mathbf{y}^{(k)}}/\partial\eta_1$. In the next section, we illustrate the use of (3.14) in the case $m = 2$.

3.2.3 Details in case $m = 2$

In this section we give further technical details for the case $m = 2$. Here the number of terms in the summation (i.e. $2^m = 4$) is manageable. In practice we use the tables for $p_{\mathbf{y}}$ and $\partial p_{\mathbf{y}}/\partial\eta_1$ to construct tables for each of the components of the weight matrix, $W_{11}, W_{12}, W_{21}, W_{22}$. It is the components of the weight matrix which we interpolate. We tabulate over the range $-20 \le \eta_1, \eta_2 \le 20$ on a rectangular grid with step length 0.1, and evaluate in between the grid points using bilinear interpolation on each of the entries in the matrix separately. The function `interp.surface` in the R package `fields` (Furrer, Nychka and Sain, 2010) is used to perform the interpolation.

Let us note the simplified forms of the integrals for the outcome $\mathbf{y} = (1,1)^T$. First, we have that

$$p_{11}(\boldsymbol{\eta}) = \int_{-\infty}^{\infty} h(\eta_1 + u)h(\eta_2 + u)\,\phi_{\sigma^2}(u)\,du\,, \tag{3.15}$$

where ϕ_{σ^2} is the density of a $N(0,\sigma^2)$ random variable. Secondly, the derivative $\partial p_{11}/\partial\eta_1$ is

$$\frac{\partial p_{11}}{\partial\eta_1}(\boldsymbol{\eta}) = \int_{-\infty}^{\infty} h'(\eta_1 + u)h(\eta_2 + u)\,\phi_{\sigma^2}(u)\,du\,. \tag{3.16}$$

When $m = 2$ it is adequate to tabulate p_{11} and $\partial p_{11}/\partial\eta_1$, together with the expectation function, $e : \mathbb{R} \to [0,1]$, which is defined by

$$e(\eta) = \int_{-\infty}^{\infty} h(\eta + u)\,\phi_{\sigma^2}(u)\,du\,.$$

This is the expected value of a single response with predictor having fixed part η, or equivalently the probability that this response is 1. Then, as $p_{\mathbf{y}}$ is the probability of outcome $\mathbf{y}$,

$$\begin{aligned}p_{11} + p_{10} &= P(y_1 = 1) = e(\eta_1)\\ p_{01} + p_{11} &= P(y_2 = 1) = e(\eta_2)\,.\end{aligned}$$

So that, rearranging,

$$p_{10} = e(\eta_1) - p_{11}(\eta_1, \eta_2) \tag{3.17}$$

$$p_{01} = e(\eta_2) - p_{11}(\eta_1, \eta_2)\,, \tag{3.18}$$

whereby we obtain tables for p_{10} and p_{01} from those for p_{11} and e. Also we have that

$$p_{00} + p_{10} + p_{01} + p_{11} = 1\,,$$

therefore we can obtain a table for p_{00} via

$$p_{00} = 1 - e(\eta_1) - e(\eta_2) + p_{11}\,. \tag{3.19}$$

As discussed in Section 3.2.2, we can use the table for $\partial p_{11}/\partial \eta_1$ to give a table for $\partial p_{11}/\partial \eta_2$ since

$$\frac{\partial p_{11}}{\partial \eta_2}(\eta_1, \eta_2) = \frac{\partial p_{11}}{\partial \eta_1}(\eta_2, \eta_1)\,.$$

The remaining partial derivatives are computed by differentiating (3.17)–(3.19) with respect to η_1 and η_2. This gives the following set of equations, which can be used to construct tables for the remaining partial derivatives given the tables for e and $\partial p_{11}/\partial \eta_1$:

$$\begin{aligned}
\frac{\partial p_{10}}{\partial \eta_1} &= e'(\eta_1) - \frac{\partial p_{11}}{\partial \eta_1} & \frac{\partial p_{10}}{\partial \eta_2} &= -\frac{\partial p_{11}}{\partial \eta_2}\\
\frac{\partial p_{01}}{\partial \eta_1} &= -\frac{\partial p_{11}}{\partial \eta_1} & \frac{\partial p_{01}}{\partial \eta_2} &= e'(\eta_2) - \frac{\partial p_{11}}{\partial \eta_2}\\
\frac{\partial p_{00}}{\partial \eta_1} &= -e'(\eta_1) + \frac{\partial p_{11}}{\partial \eta_1} & \frac{\partial p_{00}}{\partial \eta_2} &= -e'(\eta_2) + \frac{\partial p_{11}}{\partial \eta_2}\,.
\end{aligned}$$

Finally, the complete set of tables for $p_{\mathbf{y}}$, $\partial p_{\mathbf{y}}/\partial \eta_1$ and $\partial p_{\mathbf{y}}/\partial \eta_2$, $\mathbf{y} \in \{0, 1\}^2$, can be used to construct a table for the component functions of the weight matrix, via (3.12).

3.3 Adjusted MQL

We will see in Section 3.4 that the MQL approximation is not able to find quite the right treatments for the optimal design in certain examples. We suggest that this is because the approximation does not match the marginal mean of the response accurately enough, and use this to propose a simple modification which appears to improve the designs in the case of the random intercept model. This improved approximation is referred to as adjusted MQL.

Breslow and Clayton (1993) derive the MQL approximation by considering the quasi-likelihood equations for dependent data (McCullagh and Nelder, 1989, Section 9.3). Quasi-likelihood estimation requires one to specify only the mean and variance of a model, and to form the MQL estimating equations the authors use a crude approximation to these quantities. Specifically, the approximation for the mean is

$$E(y_{ij}) \approx h(\mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta}) \, ,$$

and the variance approximation comes from a first order Taylor series in the variance components.

However, in the same paper another approximation for the marginal mean of the response is mentioned in the case of a binary response with logit link. It is stated that the marginal model for the mean is a GLM with 'attenuated' coefficients. Namely, translating to the notation of Section 2.1,

$$E(y_{ij}) \approx h\left(\frac{\mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta}}{\sqrt{1 + c^2 \mathbf{z}_{ij} G \mathbf{z}_{ij}^T}}\right) \, , \tag{3.20}$$

where $c = 16\sqrt{3}/(15\pi)$, and $\mathbf{z}_{ij} = \mathbf{z}(\mathbf{x}_{ij})$. In the case of the random intercept model this simplifies to

$$E(y_{ij}) \approx h\left(\frac{\mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta}}{\sqrt{1 + c^2\sigma^2}}\right) \, .$$

Therefore, heuristically, one might anticipate the designs to be improved by substituting the attenuated version of the parameters when calculating the information matrix. The proposed reason is that in this case, the approximation to the marginal mean underlying MQL is much closer to the truth, when we consider the mean as a function of the x_i. This is a heuristic only, and a rather rough one at best, however we shall see that the approximation is very accurate in practice.

In Figure 3.3 the design resulting from the use of adjusted MQL is essentially indistinguishable from the true ML designs, and this is also seen with the Bayesian designs of Section 3.4.2. It could certainly be argued that by adjusting the parameters one is likely to corrupt the covariance approximation, but it seems that the matching the second moments accurately is much less important than matching the marginal mean. For the sake of clarity we give an explicit definition of our adjusted MQL approximation below. Note that this applies for the random intercept model only.

Adjusted MQL. Given a design ξ and parameter values $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2)^T$, the adjusted MQL information matrix of ξ at $\boldsymbol{\theta}$ is given by substituting *adjusted parameter values* into the expression for the MQL information matrix. Specifically,

$$M_{\text{AMQL}}(\xi; \boldsymbol{\theta}) = M_{\text{MQL}}(\xi; \boldsymbol{\theta}_{\text{adj}}) \, ,$$

where the adjusted parameter values, $\boldsymbol{\theta}_{\text{adj}}$, are obtained by multiplying the $\boldsymbol{\beta}$ parameters by the attenuation factor from Breslow and Clayton (1993, Section 3.1), therefore

$$\boldsymbol{\theta}_{\text{adj}} = \left(\boldsymbol{\beta}^T (1 + c^2\sigma^2)^{-1/2}, \sigma^2\right)^T \, .$$

The relative AMQL D-efficiency at $\boldsymbol{\theta}$ of design ξ_1 relative to ξ_2 is

$$\text{eff}_{\text{AMQL}}(\xi_1;\xi_2,\boldsymbol{\theta}) = \left(\frac{|M_{\text{AMQL}}(\xi_1;\boldsymbol{\theta})|}{|M_{\text{AMQL}}(\xi_2;\boldsymbol{\theta})|}\right)^{1/p} \tag{3.21}$$

$$= \text{eff}_{\text{MQL}}(\xi_1;\xi_2,\boldsymbol{\theta}_{\text{adj}})\,. \tag{3.22}$$

3.4 Examples

In this section we compute some example ML designs, and compare these to the designs resulting from the MQL and PQL approximations of Chapter 2. The random effects variance will be large enough that the designs from the various approximations are quite different. Using the ML designs as a reference we see that the MQL design is more efficient than those resulting from the other approximations of Chapter 2. The reason for this is that under MQL the selected treatments are closer to the ML-optimal treatments. The Adjusted MQL approximation proposed in Section 3.3 performs better than any of the methods from Chapter 2.

3.4.1 Locally optimal designs

We calculate designs for the 2-factor logistic random intercept model with linear predictor

$$\nu(\mathbf{x}|u) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u\,. \tag{3.23}$$

The form of the predictor is the same as in Section 2.5.2, however here there are just two points per block, and the parameter values are different. Namely, $(\beta_0, \beta_1, \beta_2) = (0, 5, 10)$ and $\sigma^2 = 5$.

As the block size is less than the number of parameters, the blocks are incomplete and the design will need more than one support block in order for the parameters to be estimable. The fixed effects parameters were chosen to be large in order to investigate a situation in which the factorial design is a poor choice (compare this to the scenario in Section 2.5.2). A large value of σ^2 was selected so that we may gain an idea of the relative performance of MQL and PQL when the designs from the two methods have diverged.

The optimal designs for this problem consist of two equally weighted blocks for each approximation. The design blocks resulting from the various methods are shown in Figure 3.3, with corresponding blocks from the different designs shown on the same plot. Additionally, we include a design labelled ‘GLM’. The points for this design were obtained by calculating a four-point exact design for the GLM corresponding to the case $\sigma^2 = 0$. The allocation shown is the optimal allocation of these points to two blocks of size two, found by computing the objective function value for the three possibilities. (Note the similarity with Section 3.1).

Using the MLNI design as a reference, the D-efficiencies of the MQL, PQL and GLM designs were 91.2%, 78.2% and 67–86.7% respectively, where the range of values for the GLM design efficiency corresponds to the use of different allocations. From the size of this range we see that allocation is indeed important. The MQL design is the most efficient out of the MQL, PQL and GLM designs. The PQL design is the least efficient of these three, being worse than the optimal allocation of the GLM design points. If we inspect the points more closely (Figure 3.4), we see that the MQL points lie between the GLM points and those from the MLNI design.

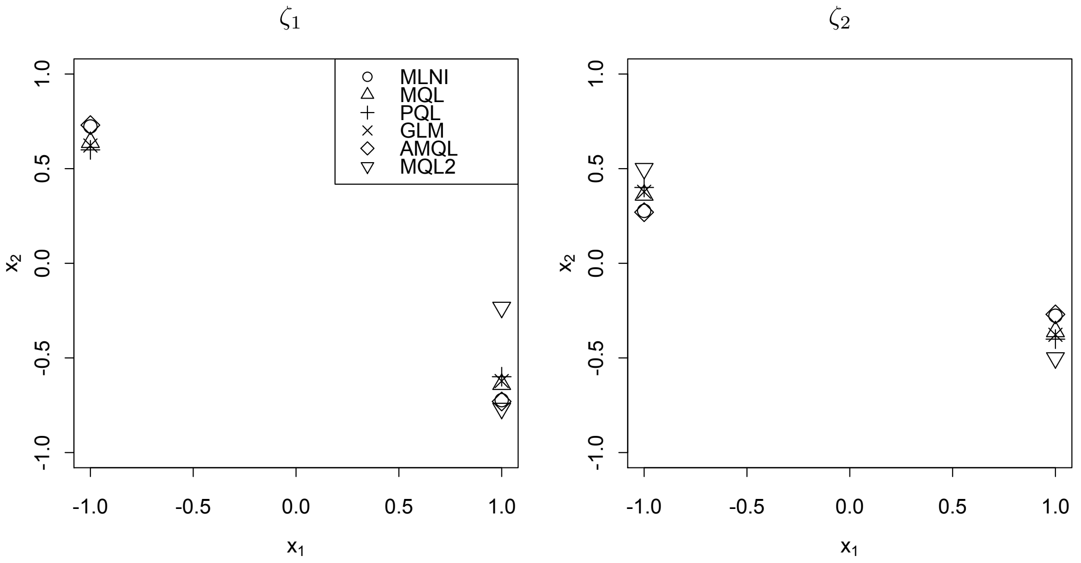

Figure 3.3: Support blocks, ζ_1 and ζ_2, in the locally D-optimal designs for model (3.23) under several approximations, approximation indicated by plotting character

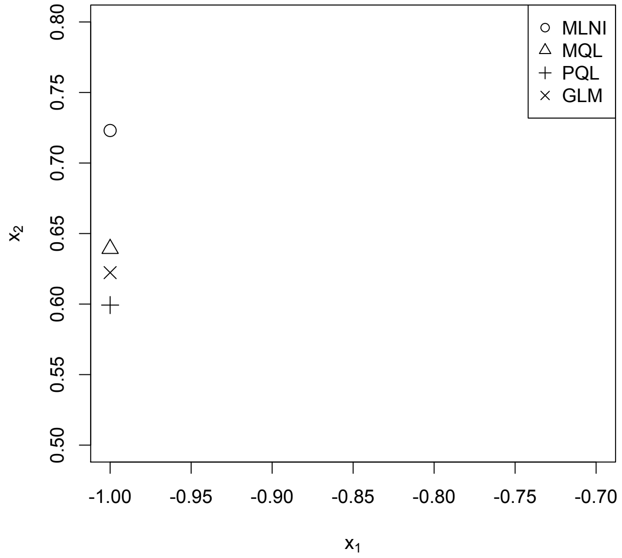

Figure 3.4: Points in the top-left quadrant of $[-1,1]^2$ from the first block of a locally D-optimal design for model (3.23) under several approximations, indicated by plotting character

Thus, the use of MQL represents a correction of the GLM design points, even if the correction is not large enough in magnitude. In contrast, the PQL points are further away from the MLNI points than the GLM points. This pattern is also seen in the 1-factor example in Chapter 4. The fact that PQL is also insensitive to the choice of allocation for large σ^2 (Section 3.1) leads us to believe that PQL is in fact a poor approximation for the purpose of producing designs for larger values of σ^2. The best approximation is AMQL, which produces a design which is virtually indistinguishable from the ML design. The MQL2 design has a point in the first block which is wildly different from that under the other approximations. Moreover in the first block, both treatments have the same value of x_1. For this particular design, it seems most likely that the optimisation has not converged to a global maximum.

3.4.2 Bayesian designs

We calculate Bayesian optimal designs for the logistic random intercept model with linear predictor

$$\nu(\mathbf{x}|u) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u\,,$$

where a priori $(\beta_0, \beta_1) = (0, 5)$, $\sigma^2 = 5$ and $\beta_2 \sim U[0, 10]$. This amounts to taking the example of Section 3.4, and introducing substantial uncertainty as to the value of β_2. The wide range of possible values of this parameter will lead to a locally optimal design being a poor choice in this example.

We choose the design ξ to optimise the value of the mean log-determinant objective function (Chaloner and Larntz, 1989)

$$I_0(\xi) = \int_{-\infty}^{\infty} \log|M(\xi;\boldsymbol{\theta})|\, f_{\boldsymbol{\theta}}(\boldsymbol{\theta})\, d\boldsymbol{\theta}\,, \tag{3.24}$$

where $f_{\boldsymbol{\theta}}$ denotes the density function of the prior on $\boldsymbol{\theta}$. In this example we use the log-determinant criterion, as opposed to the $I_{1/p}$ criterion of Firth and Hinde (1997) which is discussed in Section 2.4, because the prior density has bounded support. We evaluate the objective function (3.24) by use of a crude quadrature scheme,

$$I_0(\xi) \approx \sum_{k=0}^{10} \frac{1}{11} \log|M(\xi;\boldsymbol{\theta}_k)|\,,$$

with $\boldsymbol{\theta}_k = (0, 5, k, 5)^T$, $k = 0, \ldots, 10$. This is equivalent to approximating the uniform prior with the discrete equiprobable prior on $\{0, 1, \ldots, 10\}$. A better quadrature scheme, such as that of Gotwalt et al. (2009) could be used, however the one above is adequate for the purposes of this example. As the value of σ^2 is the same for all the $\boldsymbol{\theta}_k$, we only need precompute one lookup table (see Sections 3.2.1–3.2.2 for details of the MLNI method).

Bayesian D-optimal designs were computed under each of the different approximations, and are given in Tables 3.1–3.5. All of the Bayesian designs contain more blocks than the locally optimal design in Section 3.4, which makes sense given the degree of uncertainty in the parameters.

Locally D-optimal MLNI designs were also computed for the parameter values at each of the quadrature points $\boldsymbol{\theta}_k$. The locally optimal designs were used to compute the local D-efficiencies of the Bayesian designs, which are shown as a function of β_2 in Figure 3.5. The ordering of the performance of the approximations is clearly the same as in Section 3.4. The Bayesian MLNI and AMQL designs are essentially indistinguishable in terms of their efficiency curve. The MQL design is consistently more efficient than the PQL design. For higher values of the parameters, $\beta_2 \geq 8$, the MQL2 design slightly outperforms the MQL and PQL designs, but on average the latter two are superior to MQL2. With respect to the discrete, approximating prior the mean D-efficiencies are 80.6%, 73.5% and 66.9% for the MLNI, MQL and PQL designs respectively.

In Figure 3.6 the robustness of the optimal Bayesian MLNI design is benchmarked against some simpler comparators, namely: (i) the locally optimal MLNI design evaluated at the centroid of the parameter space i.e. $\boldsymbol{\theta} = (0, 5, 5, 10)^T$, and (ii) an optimal allocation of the 2^2 factorial design points to two equally weighted blocks, specifically $\zeta_1 = \{(-1, -1), (1, 1)\}$, $\zeta_2 = \{(-1, 1), (1, -1)\}$. We see that the Bayesian design is indeed more robust than the locally optimal design. When the parameter β_2 takes its extreme values 0 and 10, the Bayesian design has D-efficiencies of 58.1% and 65.5% respectively, compared with 46.4% and 55.6% for the locally optimal design. This greater robustness of the Bayesian design comes at the loss of a mere 2% in D-efficiency compared with the locally optimal when $\beta_2 = 5$. The factorial design performs reasonably well when β_2 is small, but extremely badly for larger values: the D-efficiency when $\beta_2 = 10$ is just 3.8%, and the D-efficiency averaged across the set of plausible values of β_2 is 31.8%.

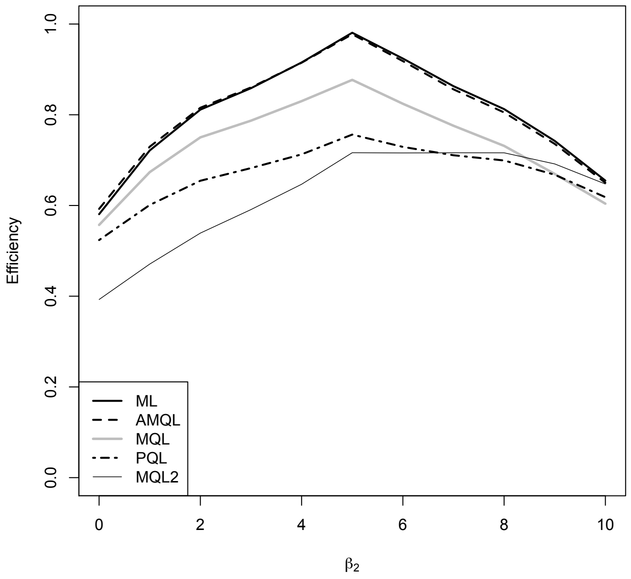

Figure 3.5: Local D-efficiencies of the Bayesian designs under each approximation

Block (i)	$\mathbf{x}_{i1}^T$	$\mathbf{x}_{i2}^T$	Weight (w_i)
1	(1.000, -0.402)	(-1.000, 0.402)	0.370
2	(-0.144, 1.000)	(0.144, -1.000)	0.227
3	(1.000, -1.000)	(-1.000, 1.000)	0.402

Table 3.1: Bayesian MLNI design

Block (i)	$\mathbf{x}_{i1}^T$	$\mathbf{x}_{i2}^T$	Weight (w_i)
1	(-0.874, 1.000)	(0.874, -1.000)	0.205
2	(-0.449, 0.049)	(0.449, -0.049)	0.328
3	(-1.000, 0.659)	(1.000, -0.659)	0.294
4	(0.232, -1.000)	(-0.232, 1.000)	0.173

Table 3.2: Bayesian MQL design

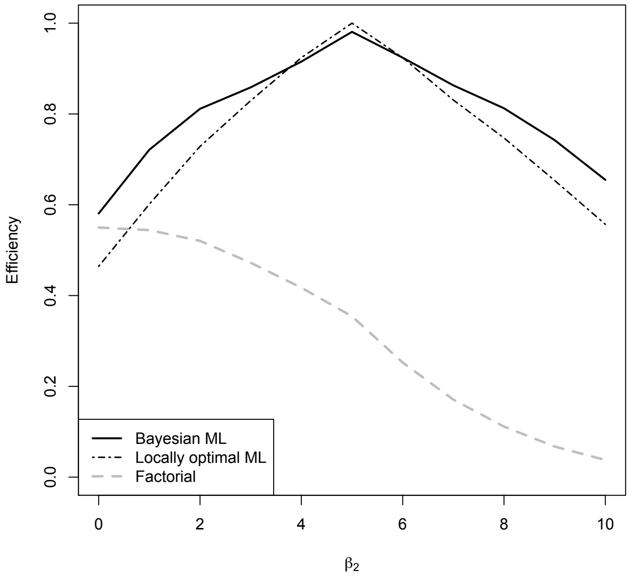

Figure 3.6: Robustness of the Bayesian MLNI design compared with simpler approaches

Block (i)	$\mathbf{x}_{i1}^T$	$\mathbf{x}_{i2}^T$	Weight (w_i)
1	(-1.000, 0.549)	(1.000, -0.592)	0.244
2	(0.213, -1.000)	(1.000, -0.499)	0.135
3	(-0.106, -0.172)	(-0.076, 1.000)	0.152
4	(0.549, -0.618)	(-0.561, 0.621)	0.379
5	(-0.422, 1.000)	(-1.000, 0.709)	0.090

Table 3.3: Bayesian PQL design

Block (i)	$\mathbf{x}_{i1}^T$	$\mathbf{x}_{i2}^T$	Weight (w_i)
1	(1.000, -0.508)	(0.261, -0.539)	0.512
2	(-1.000, 1.000)	(-1.000, 0.460)	0.254
3	(-1.000, 0.704)	(-0.316, 1.000)	0.234

Table 3.4: Bayesian MQL2 design

Block (i)	$\mathbf{x}_{i1}^T$	$\mathbf{x}_{i2}^T$	Weight (w_i)
1	(-1.000, 1.000)	(1.000, -1.000)	0.404
2	(0.129, -1.000)	(-0.128, 1.000)	0.232
3	(-1.000, 0.386)	(1.000, -0.386)	0.364

Table 3.5: Bayesian AMQL design

3.5 Linking design and analysis

In this section we make some considerations regarding the analysis of data resulting from our designs for the logistic GLMM. We perform ML estimation of the random intercept logistic model using the `R` package `glmmML` (Broström, 2011). To carry out PQL estimation we use the `glmmPQL` function in the package `BradleyTerry2` (Turner and Firth, 2010). The latter provides an option either to estimate σ^2 or to hold it fixed at a known value.

Study 1: First of all we observe that difficulties arise if we attempt to estimate all of the parameters (including σ^2) using designs with two points per block. To illustrate the problem we simulated 100 samples of $n = 100$ blocks using the optimal ML design with parameter values $(\beta_0, \beta_1, \beta_2, \sigma^2) = (0, 5, 10, 5)$. For each simulated dataset we calculated the ML estimates of the parameter values, thereby producing a sample of 100 draws from the distribution of $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ under the ML design. The resulting distributions of the ML estimators were bimodal, see for example Figure 3.7 which shows the distribution of $\hat{\beta}_1$. The range of the estimates is extremely large compared to the true parameter value. Bimodality also occurs if we use PQL estimation, and if we use the MQL or PQL designs. The issue appears to be one of parameter identification: in this case the blocks are not large enough to estimate all of the parameters (at least not with this design, which does not take account of the need to estimate σ^2). Note that a larger simulation size was not used because in this example the estimation routines crashed quite frequently, presumably owing to the difficulty of the estimation.

Study 2: However, the estimation is more satisfactory when the blocks are larger. We formed a design with a single block of size $m = 4$ from the design points of the locally D-optimal ML design from the previous paragraph, which had 2 points per block. Figure 3.8 shows the distribution of $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ based on 10,000 simulations of 100 blocks. The parameter values used were the same as before, in other words $(\beta_0, \beta_1, \beta_2, \sigma^2) = (0, 5, 10, 5)$. We see that the range of estimated values of β is vastly reduced in comparison with the case $m = 2$.

Study 3: We now compare the different designs and estimation methods, holding σ^2 fixed at its true value throughout the estimation. Figures 3.9–3.11 give Monte Carlo samples from the distributions of the parameter estimators of $\beta_0, \beta_1, \beta_2$ under ML and PQL estimation. Samples were generated using the ML, MQL and PQL designs. The true values of the parameters were set to $(\beta_0, \beta_1, \beta_2, \sigma^2) = (0, 5, 10, 5)$ and these are indicated on each plot by a vertical line. We used 10,000 repeated samples, and each sample consisted of $n = 100$ blocks in total. There are a few points of note. Firstly, when using PQL estimation, the estimators of all the parameters except β_0 are biased. Secondly, the variances of the estimators are smallest using the ML design *under both estimation methods.* In other words, using the PQL design rather than the ML design does not lead to the PQL estimator having smaller variance. The MQL design gives smaller variances than the PQL design, again under both estimation methods. In this instance, the bias under PQL estimation is somewhat reduced by using the PQL design. However, here you would not be likely to use PQL estimation in practice since ML produces nearly unbiased estimates and is not difficult to implement.

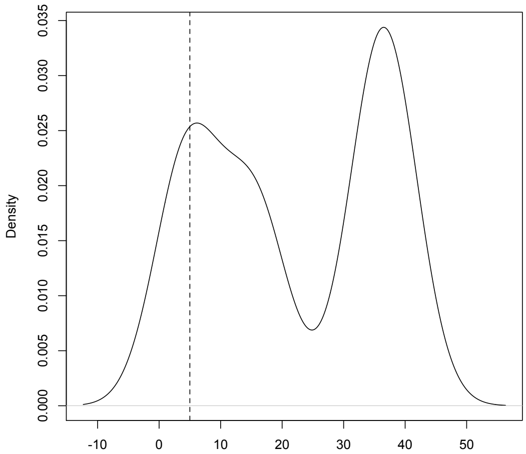

Figure 3.7: Study 1. Empirical distribution of $\hat{\beta}_1$ using the ML design with 2 points per block, based on 100 simulations of 100 blocks. Vertical line indicates the true value of β_1.

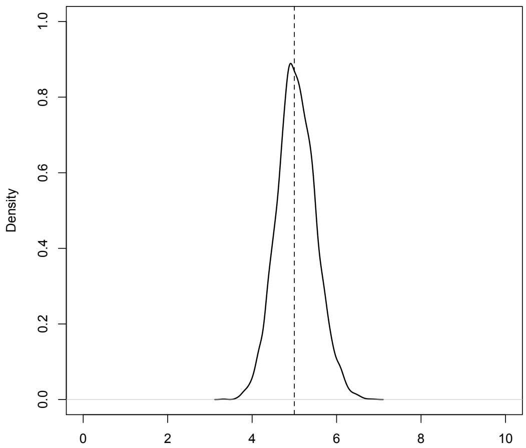

Figure 3.8: Study 2. Empirical distribution of $\hat{\beta}_1$ using the ML design with 4 points per block, based on 10,000 simulations of 100 blocks.

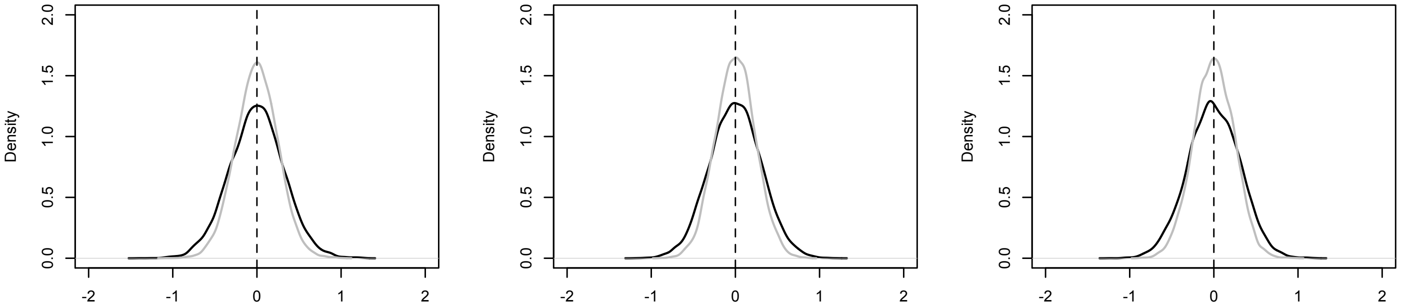

Figure 3.9: Study 3. Empirical distribution of estimates of β_0 under ML (black) and PQL (grey), using the ML, PQL and MQL designs (left, centre and right respectively).

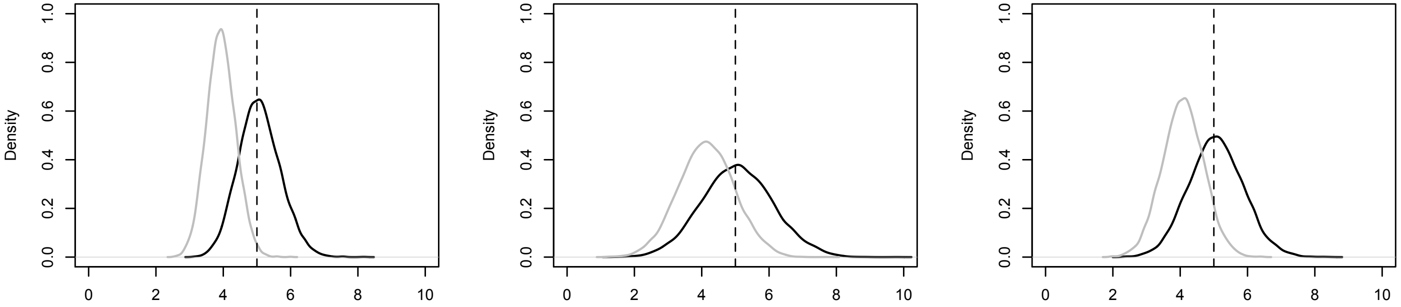

Figure 3.10: Study 3. Empirical distribution of estimates of β_1 under ML (black) and PQL (grey), using the ML, PQL and MQL designs (left, centre and right respectively).

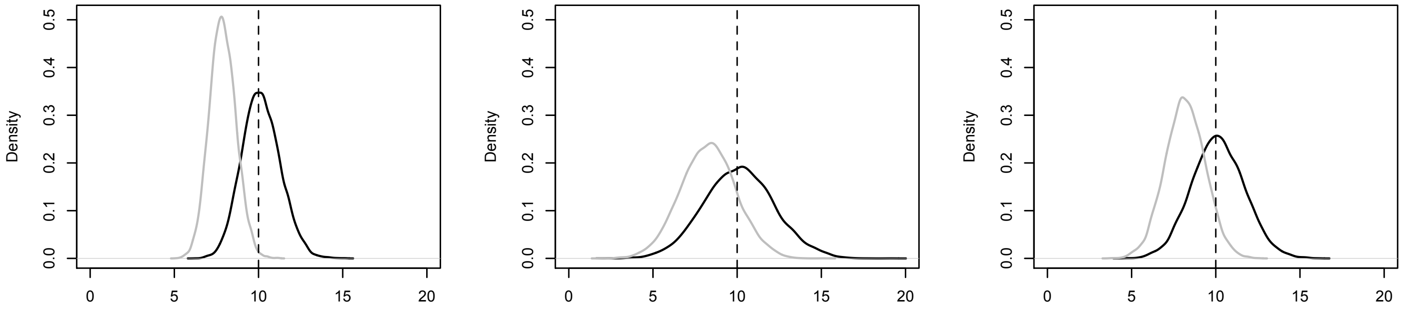

Figure 3.11: Study 3. Empirical distribution of estimates of β_2 under ML (black) and PQL (grey), using the ML, PQL and MQL designs (left, centre and right respectively).

3.6 Optimisation algorithms

In Sections 2.5 and 3.4, we performed numerical searches for optimal designs using the transformations of Atkinson et al. (2007) together with the general purpose BFGS or Nelder-Mead algorithms. In this section, we consider the use of alternative algorithms. We find that a modified co-ordinate optimisation approach is more effective.

3.6.1 Algorithm 1: Transformation

Here we give details of the transformations from Atkinson et al. (2007). These are used to convert the optimisation problem into an equivalent formulation which involves a search over an unconstrained space. The controllable treatment variables $x_1, \ldots, x_q$ are transformed according to

$$x_i = \sin z_i \,, \qquad i = 1, \ldots, q \,,$$

thus as x_i varies in $[-1, 1]$, the transformed variable z_i takes values spanning the whole of $\mathbb{R}$.

The weights w_k, $k = 1, \ldots, b$, of the design ξ are transformed in a more complicated way to account for the constraints $w_k \geq 0$, $\sum_{k=1}^{b} w_k = 1$. The transformation used is

$$\begin{aligned} w_1 &= \sin^2 \Omega_1 \\ w_2 &= \sin^2 \Omega_2 \cos^2 \Omega_1 \\ &\vdots \\ w_k &= \sin^2 \Omega_k \prod_{l=1}^{k-1} \cos^2 \Omega_l \,, \qquad 2 \leq k \leq b-1 \,, \\ &\vdots \\ w_b &= \prod_{l=1}^{b-1} \cos^2 \Omega_l \,, \end{aligned}$$

and the Ω_k are allowed to take values in the whole of $\mathbb{R}$.

General purpose algorithms are applied to find the optimal treatment values and weights on the scale of z_i and Ω_k.

3.6.2 Algorithm 2: Co-ordinate optimisation

In following sections we will make use of a co-ordinate optimisation algorithm for approximate block designs which is similar to the co-ordinate exchange algorithm employed by Meyer and Nachtsheim (1995) for exact designs. This section gives details of the simple co-ordinate optimisation. A modified version, with has additional heuristic features to cope with problems specific to approximate block designs, is described in Section 3.6.3. Throughout the optimisation, we fix the maximum number, b, of distinct support blocks allowed in a design. This means that a design can be stored in an array of dimension $b \times (mq + 1)$, in which each row corresponds to a support block of the design, where q is the number of controllable variables (factors). For a detailed plan of how the design is stored in the array, see Table 3.6.

At each step of the algorithm we select an entry (or 'co-ordinate') of the array to be the focus of our attention. Depending on the type of co-ordinate selected, we consider changes to the design in one of the following ways:

1. *Treatment variable setting.* When the selected entry belongs to one of the $\mathbf{x}_{kj}$, $1 \leq k \leq b$, $1 \leq j \leq m$, (i.e. the entry is not a weight), the selected entry is varied and all other entries in the array are held fixed.

2. *Weight, not currently equal to 1.* When the selected entry is a weight we must be careful to maintain the constraint that the weights sum to unity. This is done as follows. Suppose the current value of the weight is w_i and the proposed new value is w_i'. Then we multiply the other weights, $w_j, j \neq i$, in the design by a factor of $(1 - w_i')/(1 - w_i)$, keeping the factor settings constant.

3. *Weight, currently equal to 1.* Let w_i' be the proposed value. In this case we instead set all other weights to have the same value, $w_j' = (1 - w_i')/(b - 1)$.

In accordance with these rules we change the design by varying the selected co-ordinate, and find the value which maximises the objective function. This optimal value is then kept, together with any corresponding changes to the other parts of the design. We perform the maximisation with a general purpose one-dimensional optimisation algorithm, namely the `optimize` function in `R`.

A *pass* of the algorithm involves working through the array in 'typewriter fashion', left-to-right, top-to-bottom, performing co-ordinate optimisation steps as described above. The algorithm begins by generating random designs until a nonsingular starting design is found. It then repeatedly performs passes until a pass yields no changes to the design.

The algorithm is based on a greedy heuristic, and is prone to becoming stuck in suboptimal attractor states. As a result, multiple random initialisations are used to obtain an efficient final design.

	Unit 1	Unit 2	...	Unit m	Weight
Block 1	$\mathbf{x}_{11}^T$	$\mathbf{x}_{12}^T$	...	$\mathbf{x}_{1m}^T$	w_1
Block 2	$\mathbf{x}_{21}^T$	$\mathbf{x}_{22}^T$	...	$\mathbf{x}_{2m}^T$	w_2
$\vdots$	$\vdots$		$\ddots$		$\vdots$
Block m	$\mathbf{x}_{b1}^T$	$\mathbf{x}_{b2}^T$	...	$\mathbf{x}_{bm}^T$	w_b

Table 3.6: The structure of the array used to store approximate block designs in the algorithms. Note that $\mathbf{x}_{ij}^T$ is a row vector with entries corresponding to the settings of the q factors applied to the jth unit in the ith block.

3.6.3 Algorithm 3: Modified co-ordinate optimisation

The modified co-ordinate optimisation described in this section addresses two problems with the simple version of the algorithm given in Section 3.6.2. Firstly, the simple algorithm tends to produce designs in which there are several very similar blocks, see for example the design in

Figure 3.16. It seems that we should consider 'consolidating' these similar blocks in order to simplify the description of the design. Secondly, once a block has been given zero weight in the simple algorithm, the factor settings in that block will become stagnant, since changing their values does not affect the value of the objective function.

In order to address the problem of stagnation, we modify the basic pass of the co-ordinate optimisation to handle differently those factor settings in a block i with $w_i = 0$. Instead, given a proposed value x of such a co-ordinate we first form a modified design which assigns weight 0.01 to block i, and multiplies the remaining weights by 0.99. We then choose x to maximise the objective function value of the modified design. The result is that even zero-weighted blocks are systematically improved, and if later on they become a useful addition to the design they will be 'resurrected'.

The modified algorithm also includes a *consolidation step*, which proceeds as follows:

1. evaluate the information matrix, $M_i = M(\zeta_i, \boldsymbol{\theta})$, $i = 1, \ldots, b$, for each block of the design using the appropriate approximation.

2. for every pair $i < j$, evaluate $\Delta_{ij} = M_i - M_j$.

3. if the maximum absolute value of the entries in Δ_{ij} is less than the threshold τ, transfer the weight from block j onto block i, in other words adopt the new weights $w_i' = w_i + w_j$ and $w_j' = 0$. By default we set $\tau = 10^{-4}$.

We coalesce a pair of blocks when their information matrices are similar – in other words when they are performing a similar function for the design. This method is much neater than comparing the points in the blocks, which would require us to match points in two different blocks according to the distance between them.

The overall structure of the new algorithm is now described. Let φ denote the design objective function corresponding to the optimality criterion of interest. In the following examples, φ corresponds to local D-optimality, in other words $\varphi(\xi) = \log |M(\xi; \boldsymbol{\theta})|$.

1. Randomly generate designs until a non-singular design is found and use this to initialise the search.

2. Optimise the current design by repeatedly performing passes until a complete pass yields no changes. Call the resulting design ξ_1.

3. Attempt to consolidate ξ_1. If no consolidation is possible, we terminate the algorithm and output ξ_1. If consolidation is possible, we call the consolidated design ξ_2. Clearly ξ_2 is less efficient than ξ_1.

4. Optimise ξ_2 to form ξ_3.

5. If $\varphi(\xi_3) < \varphi(\xi_1)$, then the algorithm terminates and we output ξ_1. If $\varphi(\xi_3) = \varphi(\xi_1)$ then we stop and output ξ_3. Otherwise, if $\varphi(\xi_3) > \varphi(\xi_1)$, we instead return to Step 2 with ξ_3 as the current design.

Thus the algorithm terminates when optimisation of the consolidated design does not result in an improvement compared to the unconsolidated design.

3.7 Further examples

We now evaluate the performance of Algorithms 1, 2 and 3 by considering some test problems in which we calculate locally D-optimal MQL and PQL designs. We do not calculate AMQL designs, despite this being the best approximation. We view this omission as relatively unimportant, since the MQL objective function is simply the AMQL objective function of a problem with different parameter values. As a result, the character of the optimisation problem should not differ too much between these methods. Moreover, the MQL approximation is more general since it applies also to models other than the random intercept.

3.7.1 Two factors, first order model

Locally optimal designs were evaluated for the first order model in two factors, with linear predictor

$$\nu(\mathbf{x}|u) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u\,. \tag{3.25}$$

The assumed parameter values were

$$(\beta_0, \beta_1, \beta_2, \sigma^2)^T = (0, 1, 2, 1)^T\,,$$

and designs were evaluated with all blocks of size $m = 4$.

The optimal MQL and PQL designs constructed using Algorithm 1 each consist of a single block with weight 1. The factor settings in this block are indicated in Figure 3.12. The optimal MQL design constructed using Algorithm 2 consists of two blocks. The factor settings and weights of these blocks are given in Figure 3.13. The MQL coordinate-exchange design was slightly superior in terms of D-efficiency compared to the other two designs. The relative D-efficiency of the MQL and PQL transformation designs were 99.94% and 98.16% respectively.

The general equivalence theorem (Atkinson et al., 2007, p. 122) is often used to verify the optimality of an algorithmically-derived design. To perform this check in this example would require the evaluation of a function from the 8-dimensional space of possible support blocks. Such a function would be rather difficult to visualise and so we do not follow this approach.

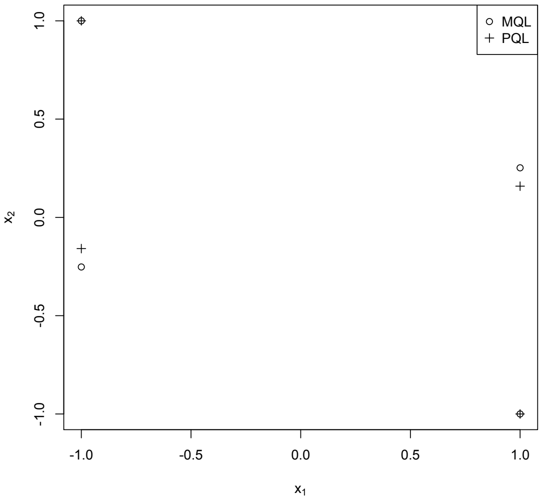

Figure 3.12: Factor settings in the MQL and PQL designs obtained using Algorithm 1 for the first-order model (3.25)

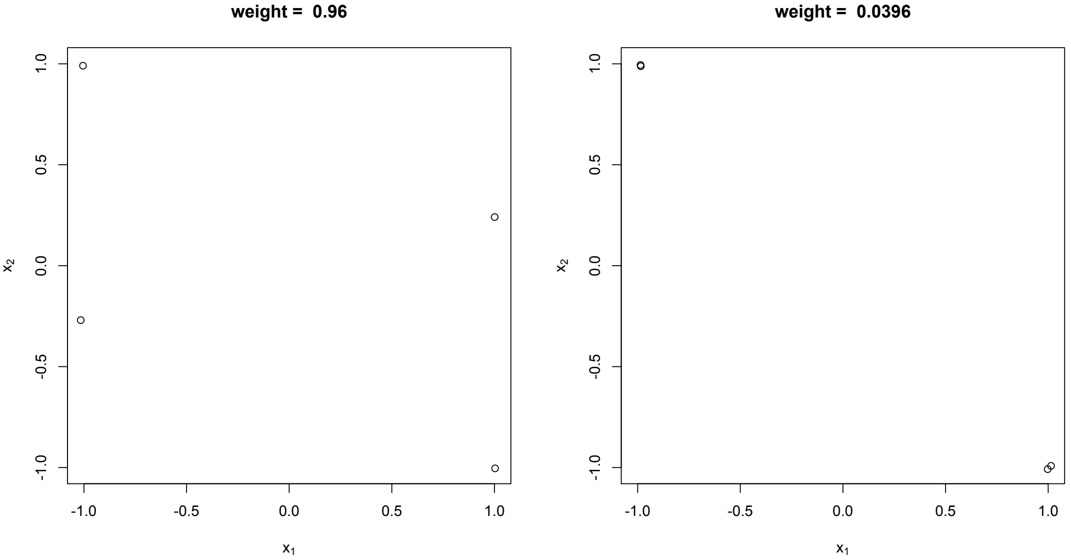

Figure 3.13: Factor settings in the MQL design obtained using Algorithm 2 for the first-order model (3.25)

3.7.2 Two factors, second order model

We consider locally optimal designs for the second order model with linear predictor

$$\nu(\mathbf{x}|u) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + u\,. \tag{3.26}$$

The assumed parameter values were $(\beta_0, \beta_1, \beta_2, \beta_{12}, \beta_{11}, \beta_{22}, \sigma^2)^T = (0, 1, 2, 0.5, 0.2, 0.2, 1)^T$. For each optimisation, the best of 10 random starts is reported.

The optimal MQL and PQL designs found by transformation each contain 3 blocks with non-negligible weight. The factor settings for these transformation designs are shown in Figures 3.14 and 3.15. MQL designs found using the co-ordinate optimisation algorithm had a larger number of blocks. When the design was allowed to include up to 10 blocks, only 8 had non-zero weight. If 5–8 blocks were allowed, all of the blocks in the resulting design had non-zero weight. For the 8-block design, see Figure 3.16. It seems that some of the blocks are very similar and could possibly be consolidated, for instance two of the blocks in the 2$^{\text{nd}}$ row.

The designs were compared using relative efficiencies calculated by complete enumeration. The PQL and MQL transformation designs designs were 97.4% and 97.9% efficient compared to the co-ordinate exchange design.

Allowing up to 10 blocks, the modified co-ordinate exchange found a design which was 4% better than the previous best (co-ordinate exchange with up to 10 blocks), and which was supported on 6 blocks. This design is shown in Figure 3.17.

3.7.3 Comparison of algorithms

The relative performance of the three algorithms; transformation, co-ordinate optimisation and modified co-ordinate optimisation, is considered here in the context of the example of Section 3.7.2. The study also helps us gain insight into the number of distinct support blocks actually required in this example.

For each method, optimal designs were computed using 10 random starts and allowing b distinct blocks, for b increasing from 2 to 10. Figure 3.18 shows the efficiency of the resulting design in each case (using the best design found overall as the reference of 100% efficiency). We see that there is essentially no improvement to be gained from allowing more than 6 blocks. It is also apparent the performance of the transformation algorithm is quite variable, and that it produces suboptimal designs. Figure 3.19 shows the number of support blocks with non-negligible weight in the final designs. From this we see that with $b \geq 6$ the simple co-ordinate optimisation will tend to include more blocks than is strictly necessary. In contrast, modified co-ordinate optimisation always produces a design containing at most 6 blocks. The transformation algorithm tends to include too few blocks, explaining the comparatively poor performance of its designs.

Figure 3.20 shows the MQL-efficiencies of the designs found using Algorithms 2 and 3 with $b \geq 4$. These efficiencies are compared with the number of blocks with non-zero weight in the final designs. We note several features of this plot. Firstly, the right hand side contains no points, meaning that none of the final designs contained 9 or 10 ‘active’ support blocks. Algorithm 3 produced three designs with six blocks. This occurs because for this algorithm allowing more than six support blocks does not increase the number of support blocks actually present in the final design. Effectively, only six blocks were needed to obtain an optimal design. There is no design on the plot with three blocks because such a design has an efficiency outside the range of the plot. Each design with $b \geq 4$ has an efficiency greater than 99.9%. Thus it is possible to use a design with just four distinct support blocks without sacrificing any appreciable efficiency. For the design to be fully efficient, at least six support blocks are required.

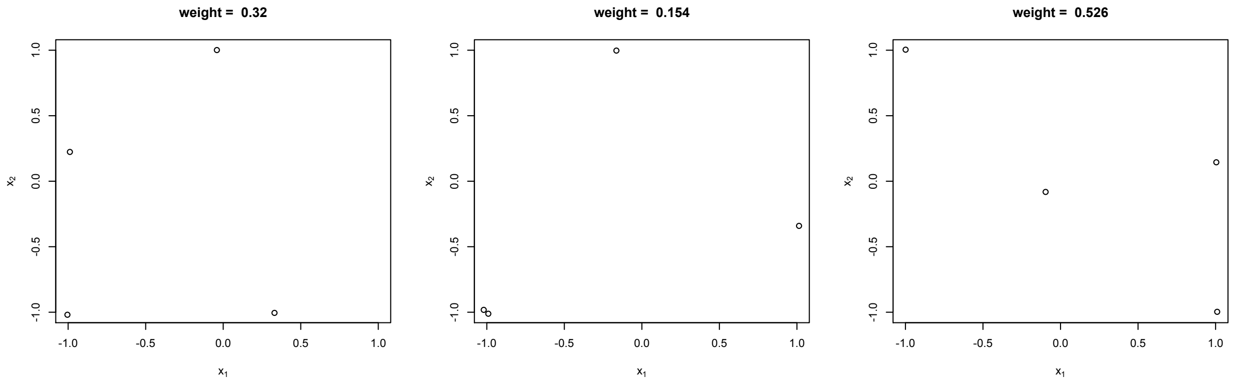

Figure 3.14: MQL design obtained using Algorithm 1 for the second-order model (3.26)

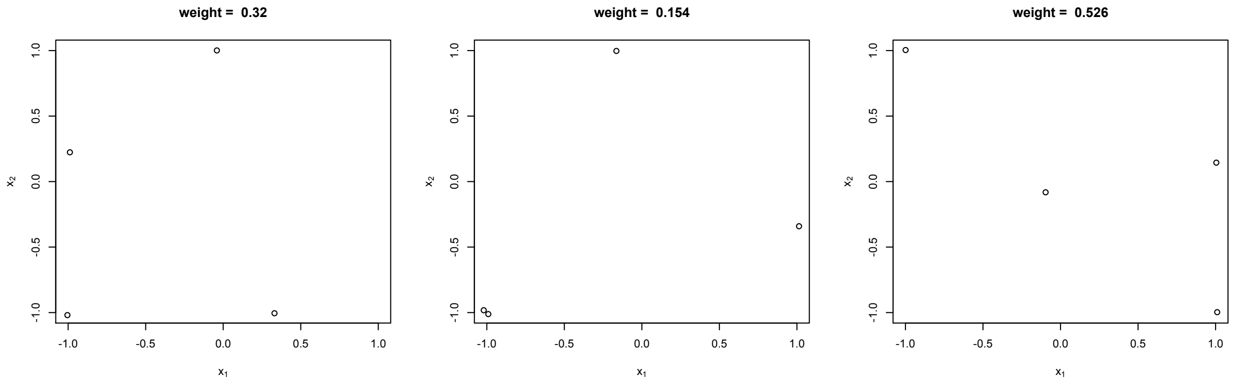

Figure 3.15: PQL design obtained using Algorithm 1 for the second-order model (3.26)

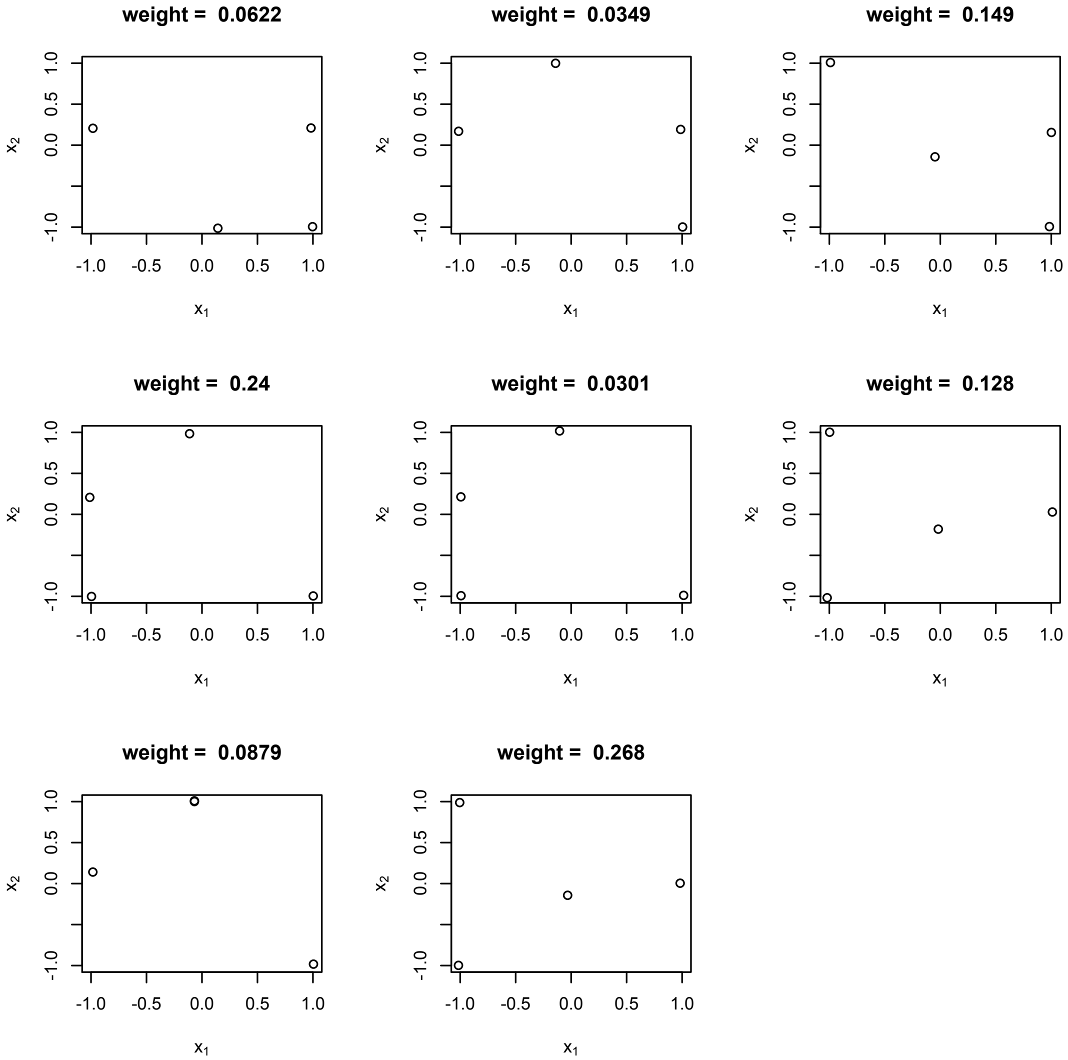

Figure 3.16: Design obtained by Algorithm 2 for the second-order model (3.26)

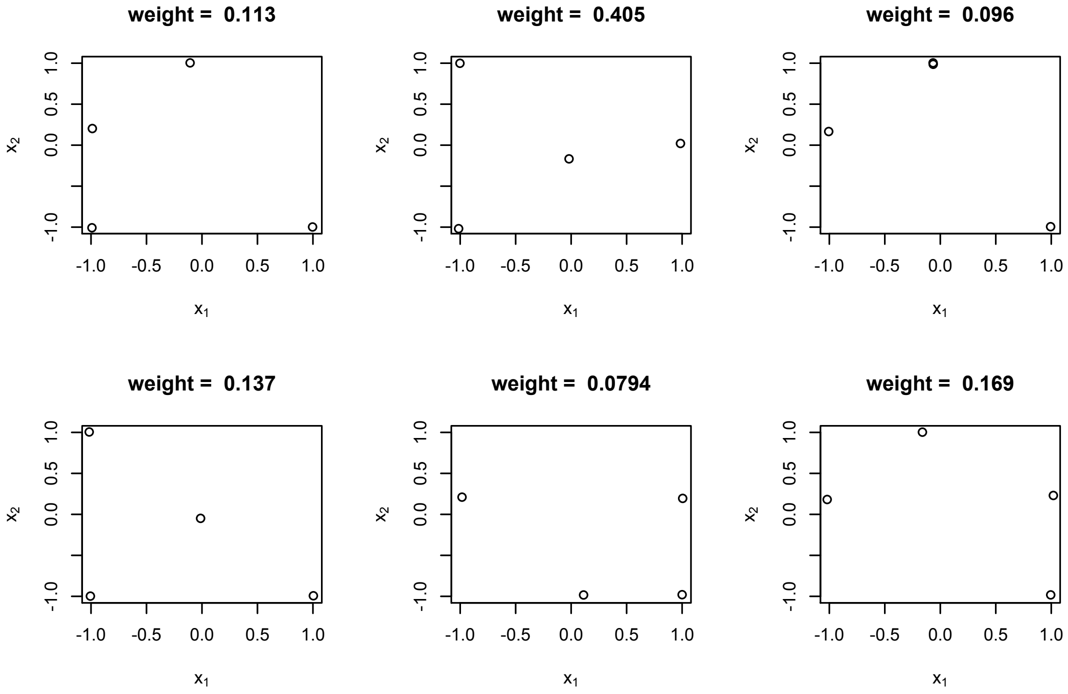

Figure 3.17: Design obtained by Algorithm 3 for the second-order model (3.26)

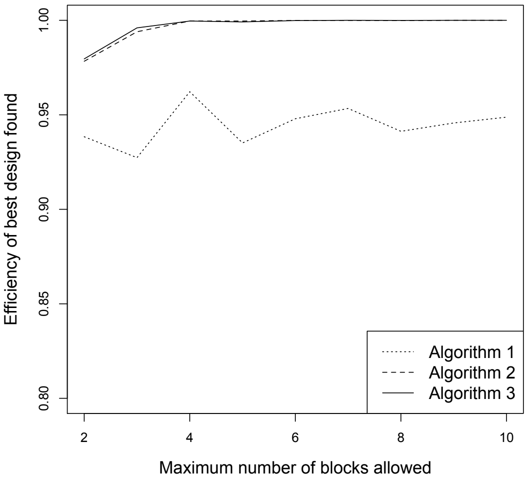

Figure 3.18: Efficiency of the designs found from the three different algorithms, with up to b blocks allowed, $b = 2, \ldots, 10$.

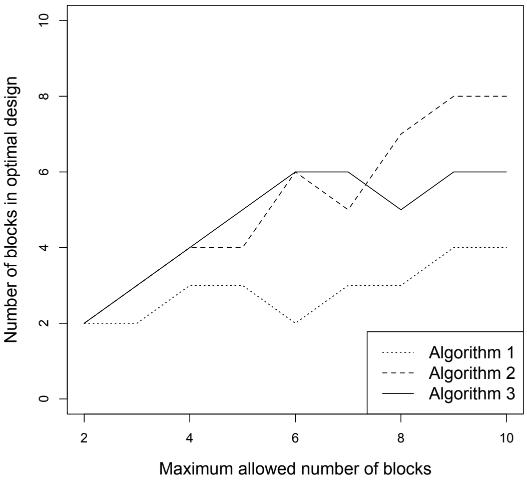

Figure 3.19: Number of blocks with non-negligible weight in the designs resulting from the three algorithms, allowing a maximum of b blocks, $b = 2, \ldots, 10$.

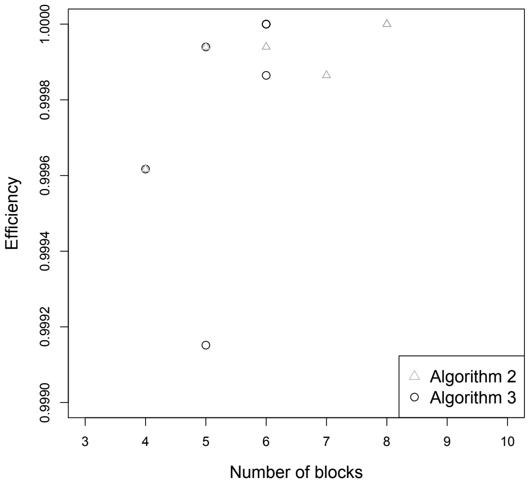

Figure 3.20: MQL efficiency of designs found versus number of blocks with positive weight

3.8 Discussion

In this chapter, one of our main objectives was to compare the performance of the designs resulting from the different approximations in Chapter 2. In order to evaluate this performance, we employed a computational methodology (MLNI) to compute ML designs in the case where there are two points per block. This comparison shows that MQL is a better approximation for design than PQL and MQL2, which have some serious deficiencies. The problems with the latter two are particularly evident when one compares different allocations for large values of the random effects variance. A new approximation, adjusted MQL, was also proposed which produced designs that were almost 100% efficient when compared to the MLNI designs.

In the future we are likely to want to calculate designs for logistic mixed models other than the random intercept, thereby allowing the effect of the x_i to vary across the blocks. The MLNI and AMQL methods are restricted to the random intercept model, but the MQL approximation can be applied immediately to the more general set up. Within the context of the random intercept model, the MLNI approach is presently restricted to two points per block. In contrast, AMQL achieves similar accuracy in the $m = 2$ case but can be used with any number of points per block.

We have seen that consideration of the block effect is worthwhile, as this impacts on the optimal treatments and the allocation of treatments to blocks. Moreover by using Bayesian criteria we were able to find designs which are substantially more robust to different values of the parameters than naïve designs such as the factorial.

3.9 Appendix: measure-theoretic results

In Section 3.2.1 we used 'differentiation under the integral sign' to express $\partial p_{\mathbf{y}}/\partial \eta_k$ as an integral. In this section, we formally state and prove the theorem which allows us to do so. This theorem is a standard result on Lebesgue integration, as is the dominated convergence theorem on which the proof rests.

Throughout this section, let (S, Σ, μ) be a measure space, where formally Σ is a σ-algebra on S and $\mu : \Sigma \to [0, \infty]$ is a measure. For details of the formalism, see e.g. Billingsley (2012). Intuitively, S is the set upon which we wish to define a measure, and Σ is the collection of subsets of S to which we can assign a meaningful value of the measure (this corresponds to an 'event' in probability theory).

Theorem 3.1 (Differentiation under the integral sign)**.** *Let U be an open subset of $\mathbb{R}$, and $f : U \times S \to \mathbb{R}$ a function satisfying*

1. *for all $t \in U$, $f_t(x) = f(t, x)$ is integrable as a function of x*
2. *for all $x \in S$, $f_x(t) = f(t, x)$ is differentiable as a function of t*
3. *there exists an integrable function g such that*

$$\left| \frac{\partial f}{\partial t}(t, x) \right| \leq g(x) \, ,$$

for all $x \in S$ and all $t \in U$.

Then, for all t, the function $d_t(x) = (\partial f/\partial t)(t, x)$ is integrable. In addition, the integral function $F : U \to \mathbb{R}$ defined by

$$F(t) = \int_S f(t, x)\, d\mu(x)$$

is differentiable and

$$\frac{dF}{dt} = \int_S \frac{\partial f}{\partial t}(t, x)\, d\mu(x)\,.$$

This is a very slight generalisation of the result given by Apostol (1974, pp. 283-4) and Williams (1991, p. 222). The proof is essentially the same, and relies on the dominated convergence theorem, which we state below. This result will also be used in Chapter 7.

Theorem 3.2 (Dominated convergence)**.** *Let $\{f_n, n \in \mathbb{N}\}$ be a sequence of real-valued measurable functions on (S, Σ, μ). Suppose that the sequence converges pointwise to a function f and that convergence is dominated by an integrable function g, in other words*

$$|f_n(x)| \leq g(x)\,,$$

for all $n \in \mathbb{N}$ and all $x \in S$. Then f is integrable and

$$\lim_{n\to\infty} \int_S f_n\, d\mu = \int_S f\, d\mu\,.$$

For details of the proof, see Billingsley (2012, p. 222).

Proof of Theorem 3.1. Let δ_n be an arbitrary sequence with $\delta_n \to 0$. Define the sequence of functions,

$$g_n(x) = \frac{f(t+\delta_n, x) - f(t, x)}{\delta_n} - \frac{\partial f}{\partial t}(t, x)\,. \tag{3.27}$$

By definition of the partial derivative $g_n(x) \to 0$ as $n \to \infty$ for all x in S. Moreover, we can rewrite this as

$$\lim_{n\to\infty} \frac{f(t+\delta_n, x) - f(t, x)}{\delta_n} = \frac{\partial f}{\partial t}(t, x)\,.$$

Thus, considered as a function of x, $\partial f/\partial t(t, x)$ is a limit of measurable functions and so is itself measurable (Billingsley, 2012, p. 194). By condition 3, it is also integrable. Also note that convergence of g_n to 0 is dominated by an integrable function: using the mean value theorem it can be seen that $|g_n| \leq 2g$. Thus, dominated convergence can be applied to show that

$$\begin{aligned}
&\lim_{n\to\infty} \frac{F(t+\delta_n) - F(t)}{\delta_n} - \int_S \frac{\partial f}{\partial t}(t, x) d\mu(x) \\
&= \lim_{n\to\infty} \int_S \left[\frac{f(t+\delta_n, x) - f(t, x)}{\delta_n} - \frac{\partial f}{\partial t}(t, x)\right] d\mu(x) \\
&= \lim_{n\to\infty} \int_S g_n(x) d\mu(x) \\
&= 0\,.
\end{aligned}$$

□

We applied Theorem 3.1 to $p_{\mathbf{y}}$, given in (3.10), in order to derive the expression (3.13) for $\partial p_{\mathbf{y}}/\partial \eta_k$. Here we demonstrate that the application of the result is valid, by checking the integrand satisfies the conditions listed in the Theorem. Here our integration variable, which was previously denoted x, is instead the vector $\mathbf{u}$ and $S = \mathbb{R}^r$. The 'parameter' of the integral, formerly t, is now η_k, which takes values in $U = \mathbb{R}$. Defining

$$f(\eta_k, \mathbf{u}) = \prod_{j=1}^{m} \Big[y_j h(\eta_j + \mathbf{z}_j^T \mathbf{u}) + (1 - y_j)\{1 - h(\eta_j + \mathbf{z}_j^T \mathbf{u})\} \Big] f_{\mathbf{u}}(\mathbf{u}) \,,$$

we have that $p_{\mathbf{y}}(\boldsymbol{\eta}, \sigma^2) = \int_{\mathbb{R}^r} f(\eta_k, \mathbf{u}) d\mathbf{u}$. We take as our measure space $\mathbb{R}^r$ with the usual Borel σ-algebra, and Lebesgue measure, and consider each condition in turn:

1. Considered as a function of $\mathbf{u}$, $f(\eta_k, \mathbf{u})$ is measurable, since it is formed from combining continuous functions of $\mathbf{u}$. Note that, as $0 \le h \le 1$, we must also have that $0 \le f(\eta_k, \mathbf{u}) \le f_{\mathbf{u}}(\mathbf{u})$. Therefore considered as a function of $\mathbf{u}$, $f(\eta_k, \mathbf{u})$ is also integrable for all η_k.

2. Considered as a function of η_k, $f(\eta_k, \mathbf{u})$ is differentiable with

$$\frac{\partial f}{\partial \eta_k} = (2y_k - 1) h'(\eta_k + \mathbf{z}_k^T \mathbf{u}) \prod_{j \neq k} \Big[y_j h(\eta_j + \mathbf{z}_j^T \mathbf{u}) + (1 - y_j)\{1 - h(\eta_j + \mathbf{z}_j^T \mathbf{u})\} \Big] f_{\mathbf{u}}(\mathbf{u}) \,.$$

3. It is the case that $0 \le h' = h(1 - h) \le 1/4$ (consider h' as a quadratic in h), and so the partial derivative is dominated by an integrable function

$$\left| \frac{\partial f}{\partial \eta_k} \right| \le (1/4) f_{\mathbf{u}}(\mathbf{u}) \,.$$

Thus we may apply the Theorem to obtain $\frac{\partial p_{\mathbf{y}}}{\partial \eta_k} = \int_{\mathbb{R}^r} \frac{\partial f}{\partial \eta_k}(\eta_k, \mathbf{u}) d\mathbf{u}$.

Part II

Dose-response experiments with unit variation

Chapter 4

Single dosing designs

4.1 Introduction

In dose-response experiments on biological organisms, there may often be substantial heterogeneity between the individuals in the study. In this scenario it may be desirable to take account of the differences between individuals by including random effects in the model, corresponding to a 'frailty' term. We will consider experiments similar to the entomological bioassay, reported by Ridout and Fenlon (1991), used to investigate the effect of virus concentration on insect mortality. In that experiment, the individuals each received a dose d of the 'treatment', and a binary response y (insect survival or death, coded 0/1) was subsequently observed. The combination of binary response and random effects means that the resulting model is related to the GLMMs of Chapters 1-3. In this chapter we consider ways of obtaining efficient designs for this problem.

One issue which arises immediately is that, compared to a vanilla GLMM experiment, there are additional restrictions on the observations we can collect. In particular, although we may be able to collect multiple observations per individual, once we observe a 1 (death) there can be no further dosings. The restriction on the set of possible outcomes impacts upon measures of estimator variability, and this must be considered in the design problem.

The feature of individuals 'dropping out' of the data upon expiring allows the model to describe the *selection effect* resulting from repeated applications of the treatment. The latter occurs since frailer individuals will tend to perish first. This selection phenomenon has been noted before in very similar models for repeated Bernoulli trials by Xue and Brookmeyer (1997), and it is also similar to the effect observed by Hougaard (1995) in models for continuous survival times with individual frailty terms.

The complication for design which is caused by the presence of the 'stopping rule' can be avoided if we restrict our attention to designs in which there is just one dosing event per individual, in other words only a single observation is collected for each insect. However, with this restriction the model parameters are approximately unidentifiable unless a prior estimate of the random effects variance, σ^2, is available.

In this chapter we focus our attention on these restricted, single-dosing designs. We have two objectives. Firstly, to consider variance-optimal designs in the case where a prior estimate of σ^2 is available. Secondly we explore the impact of misspecification of σ^2 on the estimates of the fixed effects parameters. In Chapter 5, we investigate designs with multiple dosings per individual.

For single dosing designs the model reduces to a GLMM, and so the approximations developed in Chapter 2 are applicable. These approximations are based on the MQL and PQL estimation methods described by Breslow and Clayton (1993). In Chapter 2, they allowed us to avoid the large number of numerical integrals necessary to consider the potential values of the random effects and all possible response patterns in a blocked binary response experiment. In this setup, where the model is relatively simple, we will find that the approximations are less effective than other approaches. However, this design problem is a useful testing ground for evaluating the performance of those methods. Moreover, the results confirm our findings in Chapter 2 that MQL is better than PQL at approximating the optimal treatment values. Indeed, PQL produces designs that are less efficient than the ones resulting from not taking into account the unit effect.

The single-dosing case with known σ^2 is also very similar to the situation where the model is a 1-factor binary response GLM, albeit with a rather unusual link function. Thus there is a connection between this work and that on variance-optimal designs for dose-response GLM-type models, for instance Ford et al. (1992) and Biedermann, Dette and Zhu (2006).

The experiment of Ridout and Fenlon (1991) satisfies the constraint of one dosing per individual. In their application, it was not possible to achieve the intended dose exactly, which we refer to as a *dose-error* setting. The authors' analysis addressed the presence of dose-error rather than individual variability. However, the models for these two phenomena are similar and we explore the connection between designs in the two cases. Design when the intended factor level is not achieved exactly has been discussed in the case of linear models with complicated predictor structures by Donev (2004). For dose-error models with binary response, designs have been calculated by Tang and Bacon-Shone (1992). A probit link function was chosen by those authors because the marginal mean of the response is then tractable. With a logistic link, this property fails. In this chapter we calculate designs which take into account individual variation, in other words designs for the logistic random intercept model.

A sufficiently precise prior estimate of σ^2 may be available, for instance from prior scientific knowledge. If no such estimate is available, or if our main interest is in understanding selection effects, then we must consider using designs with multiple dosings per individual. In Chapter 5, we consider the application of optimal design theory to multiple-dosing designs within a restricted class. Namely, we constrain dosings applied to the same individual to use the same dose level. The ideas of complete enumeration and precomputation, developed in Chapters 2 and 3 both prove helpful in making the calculation of designs a reality.

As the primary model in this chapter is nonlinear, the optimal design may depend on the unknown values of the model parameters. This issue of parameter dependence can be addressed by using a Bayesian design approach, which codifies uncertainty about the parameter values using a prior distribution. For thorough examples of the use of Bayesian designs for logistic regression models (without random effects), see Chaloner and Larntz (1989) or Woods et al. (2006).

4.2 Models

We begin by stating the form of the individual variation model for a general design with multiple dosings per individual. Let us denote by y_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, t_i$, the response on the jth dosing of the ith individual, which takes value 1 if the individual dies, or 0 if it survives. Also let $x_{ij} = \log d_{ij}$ be the log-dose administered on this occasion.

We assume that there are independent random effects (or 'frailty' terms), $u_i \sim N(0, \sigma^2)$, corresponding to the ith individual. Conditional upon u_i, and also provided the individual has not already died, the response of individual i on the jth occasion (or 'dosing') follows a random intercept logistic model for the probability of death, i.e.

$$\operatorname{logit} E(y_{ij}) = \beta_0 + \beta_1 x_{ij} + u_i \,, \tag{4.1}$$

where logit : $q \mapsto \log\{q/(1-q)\}$, $q \in (0,1)$. Let $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$. Then we refer to $\eta = \eta(x; \boldsymbol{\beta}) = \beta_0 + \beta_1 x$ as the *(linear) predictor*, and denote $\eta_{ij} = \eta(x_{ij}; \boldsymbol{\beta})$. The stopping rule mentioned in the introduction is that if $y_{ij} = 1$, then there are no further responses y_{ik}, $k > j$, in other words $t_i \leq T_i$, where T_i is the (random) time of the first 1 for individual i. We may stop taking observations before time T_i. For example, there may be a non-random upper limit m_i on the number of dosing events, in which case $t_i = \min\{m_i, T_i\}$. Thus in this scenario t_i is a random variable with possible values $\{1, \ldots, m_i\}$.

The number of doses survived is very much like a discrete lifetime variable, and it is interesting to consider the model from a survival analysis perspective. As noted by Xue and Brookmeyer (1997), heterogeneity of frailty leads to differential survival: since strong individuals tend to survive longer, the surviving population becomes more resilient over time. This manifests itself in a hazard function which decreases over time. Let us assume that individuals repeatedly receive a dose with linear predictor $\eta = \beta_0 + \beta_1 x$. Then

$$\begin{aligned} H(t; \eta, \sigma^2) &= P\big\{\text{die on } (t+1)\text{th dose} \,\big|\, \text{survived first } t \text{ doses}\big\} \\ &= \frac{\int_{-\infty}^{\infty} \{1 - h(\eta + u)\}^t h(\eta + u)\, \phi_{\sigma^2}(u)\, du}{\int_{-\infty}^{\infty} \{1 - h(\eta + u)\}^t\, \phi_{\sigma^2}(u)\, du} \,, \end{aligned} \tag{4.2}$$

for $\sigma^2 > 0$, with $h : \eta \mapsto 1/(1 + e^{-\eta})$ the logistic function, and ϕ_{σ^2} the density function of a $N(0, \sigma^2)$ random variable. If $\sigma^2 = 0$, then $H(t; \eta, \sigma^2) = h(\eta)$ for all t, in other words the hazard is constant. Figure 4.1 shows the change in hazard over time with $\eta = 0$ for a few different values of σ^2. We see the impact of the selection effect: for instance, if $\sigma^2 = 5.01$ and $t = 10$, only a small percentage of the remaining population will be killed following another dosing. This kind of model might potentially be useful in examples where we are interested in understanding whether further applications of, say, a particular pesticide, or antibiotic, are to provide a benefit – or whether the dose needs to be increased.

Turning around the definition of the response, so that 1 is the event 'patient is cured', and 0 is the event 'patient not cured', this model could be useful in clinical trials where some patients are intrinsically more difficult to cure than others. Clearly, once a patient is cured we would not subject them to further doses and so they would drop out of the study.

There are many examples of this discrete response bioassay setup with $t_i = 1$ for all i, in

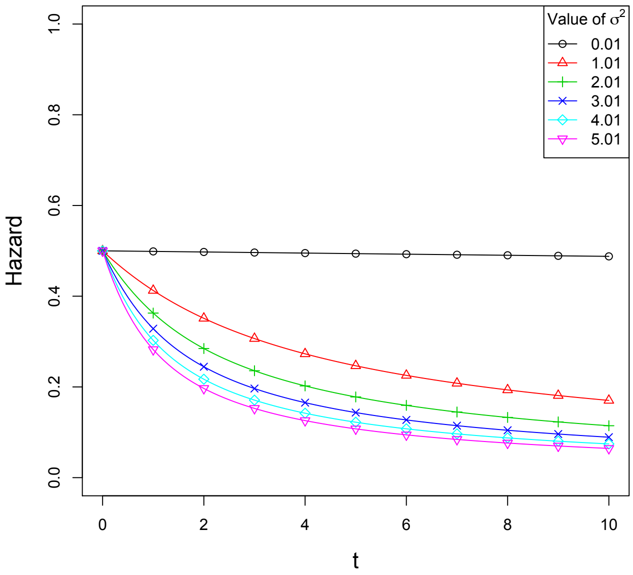

Figure 4.1: Change in the hazard, (4.2), over time for several values of σ^2

other words where there is a single dosing per individual. Usually in this case, it is assumed that $\sigma^2 = 0$. In the entomology experiment of Ridout and Fenlon (1991), x_{i1} corresponds to the log-concentration of virus which is given to the insects. The use of the log-dose in the predictor ensures a zero probability of death at zero dose, and is well established in the applied literature, see for example Smits and Vlak (1988) or Ridout and Fenlon (1991). More sophisticated analyses such as the generalised one-hit model (Ridout, Fenlon and Hughes, 1993) are better able to model the behaviour at low doses, by allowing for control mortality, in other words insect death not due to the virus.

4.2.1 Single dosing case

The remainder of this chapter focusses on designs in which each individual receives a dose on only one occasion, in other words $t_i = 1$ for all i. We refer to such designs as *single-dosing designs*.

There are several rationales for contemplating this restriction. Firstly, if $t_i = 1$ then the stopping rule is irrelevant: there can be no deaths before the first dosing, and a death on the first observation does not affect subsequent observations because there are none. The model in this case reduces to a GLMM. As a result, the design problem is simplified and we are also able to consider the performance of the approximations from Chapters 2–3 in a simpler setting than before. Secondly, an implicit assumption of model (4.1) is that individuals are not weakened by previous doses, and that the does do not accumulate. This may be true if we are careful and allow adequate recovery time for the individuals. However, if we are unsure we can avoid these assumptions by considering the case where $t_i = 1$. The downside of this is that single-dosing designs do not permit satisfactory estimation of all the parameters in the model. We show in Section 4.8 that for reasonable sample sizes, estimation of the full parameter vector is very imprecise.

Let us simplify our notation for the single-dosing case, denoting y_{i1} instead by y_i. Observe that (4.1) can be written in the notation of Section 2.1 as

$$\operatorname{logit} E(y_i \mid u_i) = \mathbf{f}^T(x_i)\boldsymbol{\beta} + u_i\,, \quad \mathbf{f}(x_i) = \begin{pmatrix} 1 \\ x_i \end{pmatrix}\,, \; \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}\,. \tag{4.3}$$

We shall denote the entire vector of model parameters as $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)^T$.

The marginal mean of the response under model (4.1) is not analytically tractable, however we have the following approximation

$$E(y_i) \approx \operatorname{expit}\left(\frac{\beta_0 + \beta_1 x_i}{\sqrt{1 + c^2\sigma^2}}\right)\,, \tag{4.4}$$

where expit is the logistic function, in other words the inverse of logit, and the constant $c = 16\sqrt{3}/(15\pi)$. Thus, the marginal mean follows approximately a logistic model with coefficients attenuated by a factor that depends on the degree of individual variation. The approximation (4.4) comes from Zeger, Liang and Albert (1988), via Breslow and Clayton (1993), and can be justified using a probit approximation to the logistic model along the lines of Demidenko (2004, pp. 335–8).

In Section 4.8, we use the approximation (4.4) to suggest a (non-neighbouring) pair of parameter vectors, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, such that a very large sample is required to distinguish between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ reliably. Therefore a prior estimate of σ^2 must be available from a separate experiment in order for the parameters of (4.1) to be estimated. If one is available, then we estimate the remaining parameters, β_0 and β_1 via the following estimation procedure.

Estimation Procedure 1. *Hold the random intercept variance, σ^2, fixed at the prior estimate, and maximise the likelihood for model (4.1) with respect to β_0 and β_1.*

In practice, an estimate of σ^2 might be available from an experiment in which there are multiple trials per individual. If these trials begin with fairly low doses then it is likely most insects will survive to receive a second treatment, enabling further observations to be collected. Moreover, with low doses the potential for weakening might be reduced. Clearly such an experiment would also give information about the other parameters, but it is not clear how to factor in this information without resorting to a Bayesian analysis.

In Section 4.6, we study the robustness of Estimation Procedure 1 with respect to misspecification of σ^2 and find that only an imprecise estimate is needed.

4.2.2 Dose-error model

We now consider the dose-error model analogous to that of Ridout and Fenlon (1991), in the single-dosing case $t_i = 1$ only. Under this model, conditioned on the received dose the response follows a logistic model, namely

$$\operatorname{logit} E(y_i \mid \epsilon_i) = \beta_0 + \beta_1(x_i + \epsilon_i)\,, \tag{4.5}$$

where the ϵ_i are independent $N(0, \sigma_\epsilon^2)$ random variables representing the deviation from the intended (log-)concentration, and $\beta_0, \beta_1 \in \mathbb{R}$, $\sigma_\epsilon^2 > 0$ are model parameters. This is identical to the model considered by Burr (1988) except that the link function here is logistic rather than probit.

The model parameters in (4.5) are also (approximately) unidentifiable unless a prior estimate of σ_ϵ^2 is available. Supposing an estimate is available, we might estimate the remaining parameters via the following procedure.

Estimation Procedure 2. *Hold the measurement error variance, σ_ϵ^2, fixed and maximise the likelihood for model (4.5) with respect to β_0 and β_1.*

This technique is similar that discussed for the logistic random intercept problem.

There are potentially simple experiments which can provide an independent estimate of σ_ϵ^2 (this was the case in Ridout and Fenlon, 1991). For example, one might repeatedly attempt to deliver a specific dose, and precisely measure the received doses using some more complicated apparatus. It would then be straightforward to estimate σ_ϵ^2 independently of the other parameters.

Clearly the dose-error model (4.5) is a reparameterisation of the random intercept model (4.1), via $\sigma^2 = \beta_1^2\sigma_\epsilon^2$. This follows from setting $u_i = \beta_1\epsilon_i$ in the original formulation. It therefore makes sense to consider the relationship between the design problems for the two models. However, note that there is a subtle difference between Estimation Procedures 1 and 2, since holding σ_ϵ^2 fixed is equivalent to holding the ratio σ^2/β_1^2 constant. The latter is clearly not the same has holding σ^2 fixed. We shall see in Appendix 4.9 that this leads to the information matrices for the two problems being different.

4.3 Optimal single-dosing designs

In this section we derive optimal designs for the unit variation model (4.1), with a single dosing per individual. We partially address parameter dependence of the optimal design using canonical forms, and the assumption that the optimal design obeys a certain symmetry property reduces the search to a one-dimensional optimisation.

We shall consider *continuous*, or *approximate*, designs. We explain this below. Suppose that there are k distinct levels among the log-doses used in the experiment, x_i, $i = 1, \ldots, n$. Without loss of generality we may reorder the data so that $x_1, \ldots, x_k$ are distinct. Moreover in this case, for all j, $k < j \le n$ there is a unique i, $1 \le i \le k$, such that $x_i = x_j$. For each i, $1 \le i \le k$, let n_i be the number of j with $x_i = x_j$. Then we have the following concise notation for the design used,

$$\xi = \left\{ \begin{array}{ccc} x_1 & \ldots & x_k \\ \lambda_1 & \ldots & \lambda_k \end{array} \right\}, \tag{4.6}$$

where $\lambda_i = n_i/n$ is the proportion of individuals assigned to log-dose x_i. The λ_i defined in this way satisfy $\lambda_i > 0$, $\sum_{i=1}^k \lambda_i = 1$. It is clear also that $n\lambda_i = n_i$ must be an integer. However, when searching for optimal designs we take (4.6) as the definition of a design, and relax the assumption that $n\lambda_i$ is an integer. Such designs are 'approximate' because for finite n the design weights λ_i will usually need to be rounded before the design is implemented.

Experimental designs are typically chosen to optimise some function of the Fisher information matrix, M, of the parameters which are to be estimated. In this work we focus on D-optimal designs, which maximise the value of $\det(M)$. The importance of the information matrix arises from its role in maximum likelihood (ML) theory, where it appears as the inverse asymptotic variance covariance matrix of the ML parameter estimators (Davison, 2003, Ch.4, p.118). Thus, D-optimal designs yield variance-optimal point estimators of the parameters to be estimated.

We concentrate on designs for Estimation Procedure 1 in Section 4.2.1. With this procedure we only estimate β_0 and β_1, whilst holding σ^2 fixed at an assumed true value. Therefore we only need consider the information matrix for $\boldsymbol{\beta}$, which we refer to as $M_{\boldsymbol{\beta}}$. This matrix will nevertheless depend on the entire set of parameters, $\boldsymbol{\theta}$. For the design ξ, the matrix $M_{\boldsymbol{\beta}}$ is given by

$$M_{\boldsymbol{\beta}}(\xi;\boldsymbol{\theta}) = \sum_{i=1}^{k} \lambda_i E_{y_i} \left\{ \frac{-\partial^2 \log p(y_i|\boldsymbol{\theta}, x_i)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} \right\}, \tag{4.7}$$

where $p(y|\boldsymbol{\theta}, x)$ is probability of obtaining response $y \in \{0,1\}$ for an individual who receives log-dose x, assuming parameter values $\boldsymbol{\theta}$. Explicitly, for $y \in \{0,1\}$, $x \in \mathbb{R}$, and $\boldsymbol{\theta} \in \mathbb{R}^2 \times [0,\infty)$ the likelihood $p(y|\boldsymbol{\theta}, x)$ is given by

$$p(y|\boldsymbol{\theta}, x) = \int_{-\infty}^{\infty} h(\beta_0 + \beta_1 x + u)^y \left\{1 - h(\beta_0 + \beta_1 x + u)\right\}^{1-y} \phi_{\sigma_u^2}(u)\, du\,, \tag{4.8}$$

where h is the logistic function, and $\phi_{\sigma_u^2}$ is the density function of a $N(0,\sigma_u^2)$ random variable. The integral in (4.8) is not tractable, therefore (4.7) can be quite costly to evaluate. This leads us to consider approximating the information matrix $M_{\boldsymbol{\beta}}$.

Breslow and Clayton (1993) and Goldstein and Rasbash (1996) discussed approximate methods for the estimation of generalised linear mixed models. The methods are referred to as marginal quasi-likelihood (MQL) and penalised quasi-likelihood (PQL), and can also be used to derive cheap approximations, $M_{\boldsymbol{\beta}}^{(a)}(\xi;\boldsymbol{\beta})$, $a =$ MQL, PQL, to $M_{\boldsymbol{\beta}}(\xi;\boldsymbol{\theta})$. For full details of the derivation of these approximations, see Chapter 2. For details of their implementation in this setting, see Section 4.4. The expression (4.7) can also be approximated directly by evaluating the integrals using numerical quadrature, although this is more costly than using MQL or PQL. Details are of this are also presented in Section 4.4. We refer to the information matrix obtained by numerical integration simply as the ML information matrix, $M_{\boldsymbol{\beta}}^{(\mathrm{ML})}$. Substituting the approximations in place of the true information matrix allows us to produce reasonably D-efficient experimental designs quickly by finding $\xi^{(a)}$ which maximises $\det(M_{\boldsymbol{\beta}}^{(a)}(\xi,\boldsymbol{\beta}))$, $a =$MQL, PQL, ML. We refer to $\xi^{(\mathrm{MQL})}$ as the 'MQL design', and so on.

Before proceeding, let us note from (4.8) that $p(y|\boldsymbol{\theta}, x)$ depends on $\boldsymbol{\theta}$ and x only through the linear predictor, $\eta = \beta_0 + \beta_1 x$. We stress this point notationally by defining, for $y = 0,1$, the function $p_y : \mathbb{R} \times [0,\infty) \to [0,1]$ via

$$p_y(\eta, \sigma^2) = p(y|\boldsymbol{\theta}, x)\,, \quad y \in \{0,1\}\,. \tag{4.9}$$

This function will be useful in Section 4.4.1 when considering in more detail the partial derivatives of $\log p$ with respect to $\boldsymbol{\beta}$.

4.3.1 Canonical forms

In this section we use a canonical form to eliminate the dependence of the optimisation problem upon $\boldsymbol{\beta}$. For a more thorough discussion on the use of canonical forms in the context of design for generalised linear models, see Atkinson and Haines (1996).

The ML information matrix, and the MQL and PQL approximate information matrices, for the design ξ can all be written in the form

$$M_{\boldsymbol{\beta}}^{(a)}(\xi;\boldsymbol{\theta}) = \sum_{i=1}^{k} \lambda_i \mathbf{f}(x_i) W(\eta_i,\sigma^2)\mathbf{f}^T(x_i)\,, \tag{4.10}$$

where $W(\eta_i,\sigma^2)$ is a scalar function of the linear predictor $\eta_i = \beta_0 + \beta_1 x_i$ and σ^2. The key point is that $M_{\boldsymbol{\beta}}$ depends on $\boldsymbol{\beta}$ only through the predictors η_i. The exact form of W depends on the approximation used, however in all cases (4.10) can be rewritten as

$$\begin{aligned} M_{\boldsymbol{\beta}}(\xi;\boldsymbol{\theta}) &= \sum_{i=1}^{k} \lambda_i\, W(\eta_i,\sigma^2)\mathbf{f}(x_i)\mathbf{f}^T(x_i) \\ &= \sum_{i=1}^{k} \lambda_i\, W(\eta_i,\sigma^2) \begin{pmatrix} 1 \\ x_i \end{pmatrix} \begin{pmatrix} 1 & x_i \end{pmatrix} \\ &= \sum_{i=1}^{k} \lambda_i\, W(\eta_i,\sigma^2) \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix}, \end{aligned} \tag{4.11}$$

where we use the notation $M_{\boldsymbol{\beta}}$ rather than $M_{\boldsymbol{\beta}}^{(a)}$ to avoid clutter. We now transform to a canonical problem using the variable

$$z_i = \beta_0 + \beta_1 x_i\,, \qquad i = 1,\ldots,k\,.$$

Note that defining

$$B = \begin{pmatrix} 1 & 0 \\ \beta_0 & \beta_1 \end{pmatrix},$$

we have that

$$\begin{pmatrix} 1 \\ z_i \end{pmatrix} = B \begin{pmatrix} 1 \\ x_i \end{pmatrix}, \qquad i = 1,\ldots,k\,, \tag{4.12}$$

and therefore, for $\beta_1 \neq 0$,

$$\mathbf{f}(x_i) = B^{-1} \begin{pmatrix} 1 \\ z_i \end{pmatrix}, \qquad i = 1,\ldots,k\,.$$

Thus the information matrix (4.11) can be written as

$$\begin{aligned} M_{\boldsymbol{\beta}}(\xi;\boldsymbol{\theta}) &= \sum_{i=1}^{k} \lambda_i\, W(z_i,\sigma^2) B^{-1} \begin{pmatrix} 1 & z_i \\ z_i & z_i^2 \end{pmatrix} B^{-T} \\ &= B^{-1} \left\{ \sum_{i=1}^{k} \lambda_i\, W(z_i,\sigma^2) \begin{pmatrix} 1 & z_i \\ z_i & z_i^2 \end{pmatrix} \right\} B^{-T}\,, \end{aligned}$$

and the determinant of M can be written as

$$\begin{aligned} |M_{\boldsymbol{\beta}}(\xi;\boldsymbol{\theta})| &= |B|^{-2} \left| \sum_{i=1}^{k} \lambda_i \, W(z_i,\sigma^2) \begin{pmatrix} 1 & z_i \\ z_i & z_i^2 \end{pmatrix} \right| \\ &= |B|^{-2} \varphi(\mathbf{z};\boldsymbol{\lambda};\sigma^2) \, . \end{aligned} \tag{4.13}$$

Above, $\mathbf{z} = (z_1, \ldots, z_k)^T$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)^T$, and φ is defined by

$$\varphi(\mathbf{z},\boldsymbol{\lambda};\sigma^2) = \left| \sum_{i=1}^{k} \lambda_i \, W(z_i,\sigma^2) \begin{pmatrix} 1 & z_i \\ z_i & z_i^2 \end{pmatrix} \right| \, . \tag{4.14}$$

Since B does not depend on $\mathbf{z}$, (4.13) is maximised with respect to $\mathbf{z}$ if and only if (4.14) is maximised. We can thus find the D-optimal design by finding $\mathbf{z}$ and $\boldsymbol{\lambda}$ to maximise $\varphi(\mathbf{z},\boldsymbol{\lambda};\sigma^2)$, and then transforming back to obtain the doses, using $x_i = \beta_1^{-1}(z_i - \beta_0)$. Hence we need now only solve an optimisation problem which depends on σ^2, and not upon $\boldsymbol{\beta}$.

Note that (4.14) is equal to the determinant of the information matrix of a transformed design for the canonical choice of parameters, $\boldsymbol{\beta}_c = (0,1)^T$. In other words,

$$\varphi(\mathbf{z},\boldsymbol{\lambda};\sigma^2) = |M_{\boldsymbol{\beta}}\big(\xi_z\,;\,\boldsymbol{\theta} = (\boldsymbol{\beta}_c\,,\,\sigma^2)^T\big)| \, ,$$

where

$$\xi_z = \begin{pmatrix} z_1 & \ldots & z_k \\ \lambda_1 & \ldots & \lambda_k \end{pmatrix} .$$

4.3.2 Symmetries

Biedermann et al. (2006) show that, for a wide variety of binary outcome dose-response models, the D-optimal design is supported on 2 distinct doses. This is one of several papers which attempt to derive more general theoretical results about the number of support points in various types of nonlinear models. Another notable example is Yang (2010). Aside from their theoretical interest, these results are helpful because they enable us to reduce the dimension of the design optimisation problem, thereby allowing more efficient and reliable numerical searches to be conducted.

In Section 4.12 we adapt the arguments of the above authors to derive similar results for models with information matrices having the structure given in (4.10). In particular we derive an analytical condition (condition I) on the weight function W, which is sufficient to guarantee that the canonical design maximising $\varphi(\mathbf{z},\boldsymbol{\lambda};\sigma^2)$ is of the simplified form

$$\xi_z = \left\{ \begin{matrix} -z & z \\ 1/2 & 1/2 \end{matrix} \right\} . \tag{4.15}$$

In words, ξ_z is supported on 2 equally weighted log-doses, and these are symmetric around 0.

We are able to establish analytically that condition I holds when W is the weight function corresponding to the MQL and PQL approximations. For the ML weight function, we have performed numerical checks which appear to confirm that condition I holds, although we have not been able to obtain an analytical proof. Therefore in all cases we restrict our numerical search for the optimal canonical design to designs of the simplified form (4.15).

It will be shown that the weight function W for each approximation satisfies

$$W(-\eta, \sigma^2) = W(\eta, \sigma^2)\,, \tag{4.16}$$

therefore the objective function (4.14) simplifies under assumption (4.15) to

$$\begin{aligned}\chi(z|\sigma^2) &= \left| W(z, \sigma^2) \begin{pmatrix} 1 & 0 \\ 0 & z^2 \end{pmatrix} \right| \\ &\propto z^2\, W(z, \sigma^2)^2\,. \end{aligned} \tag{4.17}$$

Thus, for fixed σ^2, finding the optimal design reduces to the one-dimensional problem of finding $z \in \mathbb{R}$ which maximises (4.17).

In Sections 4.10 and 4.11 we also offer alternative proofs that the optimal canonical design is of the simplified form (4.15), in the case of the MQL approximation only. One of the proofs makes use of the results of Pukelsheim (1987) and Yang et al. (2011), the other utilises the General Equivalence Theorem.

4.4 Approximations

In this Section we give the details of the computation of D-optimal designs using the different approximations. In the case of MQL, the optimal design can be derived as a function of σ^2 without having to resort to a separate optimisation for each value of σ^2.

The MQL, PQL and ML designs, $\xi^{(\mathrm{MQL})}$, $\xi^{(\mathrm{PQL})}$ and $\xi^{(\mathrm{ML})}$, are compared in Section 4.4.5. We also compare these to designs resulting from a further approximation, AGLM, which is obtained by considering a probit approximation to the logistic link function. It is found that the MQL and PQL approximations are poor in this problem, although MQL at least leads to designs which are better than those obtained by ignoring the presence of the random effect. This latter point is in agreement with the results of Chapter 3, where it was found that the MQL designs had better treatments than those under PQL.

4.4.1 Maximum likelihood

In this section we demonstrate how to calculate the weight matrix W of equation (4.10) under maximum likelihood. We also show that the weight function satisfies the property (4.16).

Recall from (4.7) that the information matrix for a general design ξ is

$$M_{\boldsymbol{\beta}}(\xi; \boldsymbol{\theta}) = \sum_{i=1}^{k} \lambda_i E_{y_i} \left\{ \frac{-\partial^2 \log p(y_i|\boldsymbol{\theta}, x_i)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right\}\,. \tag{4.18}$$

The expectation can be evaluated by considering the two possible outcomes for the response y

when log-dose x is applied,

$$E_y \left\{ \frac{-\partial^2 \log p(y|\boldsymbol{\theta}, x)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right\} = \sum_{y=0}^{1} \frac{-\partial^2 \log p(y|\boldsymbol{\theta}, x)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} p(y|\boldsymbol{\theta}, x) \,.$$

Using standard ML theory, under regularity this can be rewritten as

$$E_y \left\{ \frac{-\partial^2 \log p(y|\boldsymbol{\theta}, x)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right\} = \sum_{y=0}^{1} p(y|\boldsymbol{\theta}, x) \left(\frac{\partial \log p(y|\boldsymbol{\theta}, x)}{\partial \boldsymbol{\beta}} \right) \left(\frac{\partial \log p(y|\boldsymbol{\theta}, x)}{\partial \boldsymbol{\beta}} \right)^T ,$$

which can be further rewritten, using the chain rule on $\log p$, as

$$E_y \left\{ \frac{-\partial^2 \log p(y|\boldsymbol{\theta}, x)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right\} = \sum_{y=0}^{1} \frac{1}{p(y|\boldsymbol{\theta}, x)} \left(\frac{\partial p(y|\boldsymbol{\theta}, x)}{\partial \boldsymbol{\beta}} \right) \left(\frac{\partial p(y|\boldsymbol{\theta}, x)}{\partial \boldsymbol{\beta}} \right)^T . \tag{4.19}$$

Recalling from (4.9) that $p(y|\boldsymbol{\theta}, x) = p_y(\eta, \sigma^2)$, we have by the chain rule that

$$\frac{\partial p(y|\boldsymbol{\theta}, x)}{\partial \boldsymbol{\beta}} = \mathbf{f}(x) \frac{\partial p_y}{\partial \eta} \,. \tag{4.20}$$

Therefore, combining (4.19) and (4.20), we can express (4.18) as

$$M_{\boldsymbol{\beta}}(\xi; \boldsymbol{\theta}) = \sum_{i=1}^{k} \lambda_i \, \mathbf{f}(x_i) \left\{ \sum_{y=0}^{1} \frac{1}{p_y(\eta_i, \sigma^2)} \left[\frac{\partial p_y}{\partial \eta}(\eta_i, \sigma^2) \right]^2 \right\} \mathbf{f}^T(x_i) \,,$$

where $\frac{\partial p_y}{\partial \eta}(\eta_i, \sigma^2)$ means evaluation of the partial derivative function $\frac{\partial p_y}{\partial \eta}$ at arguments (η_i, σ^2). This is clearly in the form (4.10), therefore the weight function W under maximum likelihood is

$$W(\eta, \sigma^2) = \sum_{y=0}^{1} \frac{1}{p_y(\eta, \sigma^2)} \left(\frac{\partial p_y}{\partial \eta} \right)^2 . \tag{4.21}$$

Using the fact that $p_0 + p_1 = 1$, (4.21) can be simplified as follows

$$\begin{aligned} W &= \frac{1}{p_0} \left(\frac{\partial p_0}{\partial \eta} \right)^2 + \frac{1}{p_1} \left(\frac{\partial p_1}{\partial \eta} \right)^2 \\ &= \frac{1}{1 - p_1} \left(\frac{\partial (1 - p_1)}{\partial \eta} \right)^2 + \frac{1}{p_1} \left(\frac{\partial p_1}{\partial \eta} \right)^2 \\ &= \left\{ \frac{1}{1 - p_1} + \frac{1}{p_1} \right\} \left(\frac{\partial p_1}{\partial \eta} \right)^2 \\ &= \frac{1}{p_1(1 - p_1)} \left(\frac{\partial p_1}{\partial \eta} \right)^2 . \end{aligned} \tag{4.22}$$

Recall that p_1 and its derivative are given by

$$p_1(\eta, \sigma^2) = \int_{-\infty}^{\infty} \frac{1}{1 + e^{-(\eta + u)}} \phi_{\sigma^2}(u) \, du \,, \tag{4.23}$$

$$\begin{aligned}\frac{\partial p_1}{\partial \eta}(\eta, \sigma^2) &= \int_{-\infty}^{\infty} \frac{\partial}{\partial \eta}\left\{\frac{1}{1+e^{-(\eta+u)}}\right\} \phi_{\sigma^2}(u)\, du \\ &= \int_{-\infty}^{\infty} \frac{e^{\eta+u}}{(1+e^{\eta+u})^2}\, \phi_{\sigma^2}(u)\, du\,, \qquad (4.24)\end{aligned}$$

where ϕ_{σ^2} is the density function of a $N(0, \sigma^2)$ random variable. We evaluate W using the form (4.22), by numerical integration of (4.23) and (4.24).

To show that the weight function satisfies the symmetry property (4.16), note that; (i) $p_1(-\eta) = 1 - p_1(\eta)$, and (ii) $\partial p_1/\partial\eta$ is invariant to $\eta \to -\eta$. Therefore $p_1(1-p_1)$ is also invariant to $\eta \to -\eta$ (as the two terms in the product 'swap over') and so too is (4.22). Facts (i) and (ii) can be verified by some simple algebra using expressions (4.23) and (4.24).

4.4.2 MQL

Under MQL (see Chapter 2), the weight function in (4.10) is

$$\begin{aligned}W(\eta, \sigma^2) &= \left\{\frac{1}{\mu^{(0)}(1-\mu^{(0)})} + \sigma^2\right\}^{-1} \\ &= \left(e^{\eta} + 2 + e^{-\eta} + \sigma^2\right)^{-1}\,,\end{aligned}$$

where $\mu^{(0)} = h(\beta_0 + \beta_1 x)$ is the approximation to the conditional mean obtained by assuming that $u_i \approx 0$. It is a simple check to see that W satisfies property (4.16).

We now present a simplification of the optimisation problem which occurs for MQL only. Recall from (4.17) that finding the optimal canonical design amounts to finding z which maximises $\chi(z|\sigma^2) = z^2 W(z)^2$. Maximising χ is equivalent to minimising

$$\begin{aligned}\psi(z) &= \chi(z|\sigma^2)^{-1/2} \\ &= z^{-1} W(z)^{-1} \\ &= z^{-1}\left(e^{\eta} + 2 + e^{-\eta} + \sigma^2\right)\,. \qquad (4.25)\end{aligned}$$

It turns out to be more sensible to consider the problem in this form, as it is easier to obtain the solution over a range of σ^2 values.

Let us define a special case of the function ψ when $\sigma^2 = 0$, i.e. when the model is a GLM,

$$\psi_0(z) = z^{-1}\left(\frac{1}{\mu^{(0)}(1-\mu^{(0)})}\right)\,. \qquad (4.26)$$

Then it follows trivially from (4.25) that

$$\psi(z|\sigma^2) = \psi_0(z) + z^{-1}\sigma^2\,,$$

which when differentiated yields

$$\frac{d\psi}{dz} = \frac{d\psi_0}{dz} - \frac{\sigma^2}{z^2}\,. \qquad (4.27)$$

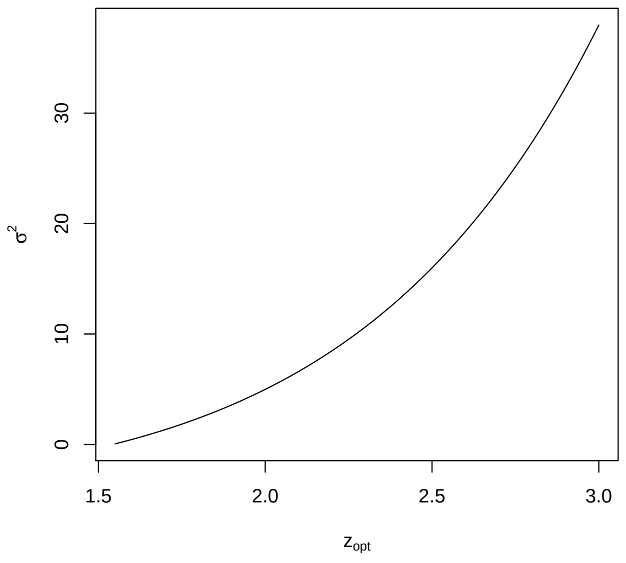

Figure 4.2: The function $g(z)$ determining the optimal MQL design, $1.543 \le z \le 3$

From (4.27), it follows that

$$\frac{d\psi}{dz} = 0 \text{ if and only if } z^2\frac{d\psi_0}{dz} = \sigma^2\,.$$

Therefore to find the optimal design ξ_z, we consider the function

$$g(z) = z^2\frac{d\psi_0}{dz}\,,$$

and for a given value of σ^2 we find z_{opt} solving $g(z) = \sigma^2$. Figure 4.2 shows the function g, therefore giving the optimal design for σ^2 up to 30. Here the derivative of ψ_0 was evaluated using numerical differentiation, however an analytical expression is available. Note that we only plot g for $z \ge 1.543$, which is the optimal value when $\sigma^2 = 0$, since g is negative for smaller values of z. For $\sigma^2 \le 5$, the optimal design is given approximately by

$$z^* \approx 1.543 + (1/10)\sigma^2\,,$$

which can be seen by reference to Figure 4.3.

4.4.3 PQL

Under PQL (for details see Chapter 2), the weight function in (4.10) is

$$W(\eta, \sigma^2) = \left\{2 + 2e^{\sigma^2/2}\cosh(\eta) + \sigma^2\right\}^{-1}\,.$$

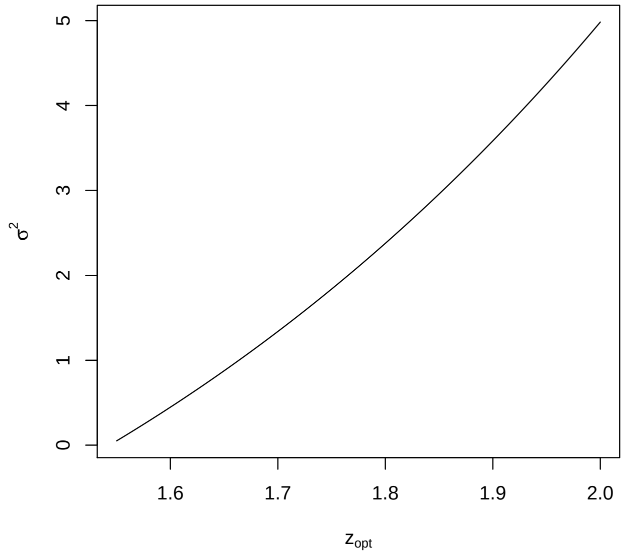

Figure 4.3: The function $g(z)$ determining the optimal MQL design, $1.543 \leq z \leq 2$

It is clear that the weight function satisfies property (4.16) since cosh is an even function. No appreciable simplification of the optimisation problem is possible.

4.4.4 Adjusted GLM

We shall also consider designs derived from a probit approximation to the logistic link. We refer to such designs as Adjusted GLM (AGLM) designs. This approximation makes the integrals analytically tractable, and leads to the attenuation formula in Breslow and Clayton (1993, section 3.1). Namely, the marginal mean is approximately

$$E(y_i) \approx \operatorname{expit}\left(\frac{\beta_0 + \beta_1 x_i}{\sqrt{1 + c^2\sigma^2}}\right) ,$$

where $c = 16\sqrt{3}/(15\pi)$. In other words the marginal model is approximately also a logistic model whose coefficients are *attenuated* by a factor $(1 + c^2\sigma^2)^{-1/2}$. Therefore to form the probit approximation to the optimal design we find the optimal design for a GLM with attenuated parameters $(\beta_0', \beta_1') = (\beta_0/\sqrt{1 + c^2\sigma^2}, \beta_1/\sqrt{1 + c^2\sigma^2})$. This results in the support points

$$\pm z_{\text{can.}} \frac{\sqrt{1 + c^2\sigma^2}}{\beta_1} - \frac{\beta_0}{\beta_1} ,$$

where $z_{\text{can.}}$ is the positive support point of the optimal canonical design for the GLM, in other words for the model with parameters $(\beta_0, \beta_1, \sigma^2) = (0, 1, 0)$.

The idea behind this approximation is quite similar to that underlying AMQL in Section 3.3. However, for AMQL finds an MQL design with adjusted parameters, rather than a GLM design. The reason it is sufficient here to adjust a GLM design is that there are no covariances

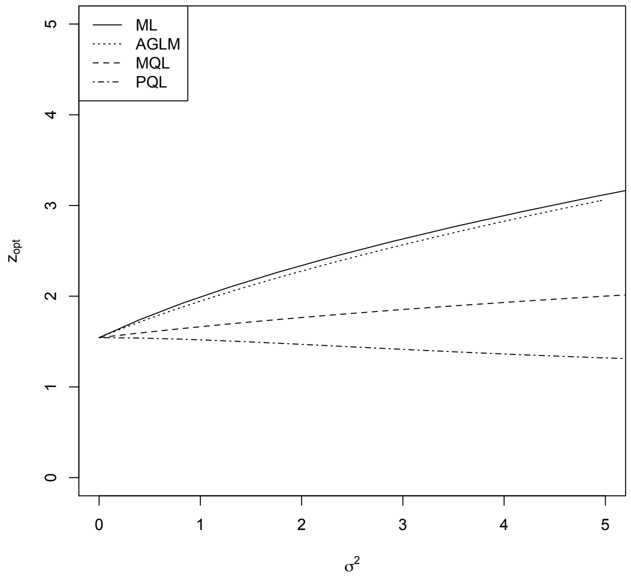

Figure 4.4: Locally optimal canonical designs under ML, MQL, PQL and AGLM, $\sigma^2 \leq 5$

to consider due to there being only one observation per individual.

4.4.5 Comparison of locally optimal designs

Figure 4.4 shows the optimal $z > 0$ for $\sigma^2 \leq 5$ under ML, MQL, PQL and AGLM approximations. Figure 4.5 shows the optimal $z > 0$ under ML and AGLM for values of σ^2 up to 100.

The D-efficiency of an arbitrary design ξ is

$$D\text{-eff}(\xi; \boldsymbol{\theta}) = \left\{ \frac{|M_{\boldsymbol{\beta}}(\xi; \boldsymbol{\theta})|}{|M_{\boldsymbol{\beta}}(\xi^*; \boldsymbol{\theta})|} \right\}^{1/2} ,$$

where ξ^* is the locally D-optimal design (under ML). Figures 4.6 and 4.7 show the efficiency of the MQL and AGLM designs as a function of σ^2. Note in particular that the efficiency of the AGLM design is extremely close to 1. Note that the PQL design points are heading in the opposite direction to the others, and therefore for $\sigma^2 > 0$ the PQL design will be less efficient than the GLM design. This is noteworthy given that the GLM design does not acknowledge the presence of individual variation, whereas the PQL design attempts to do so.

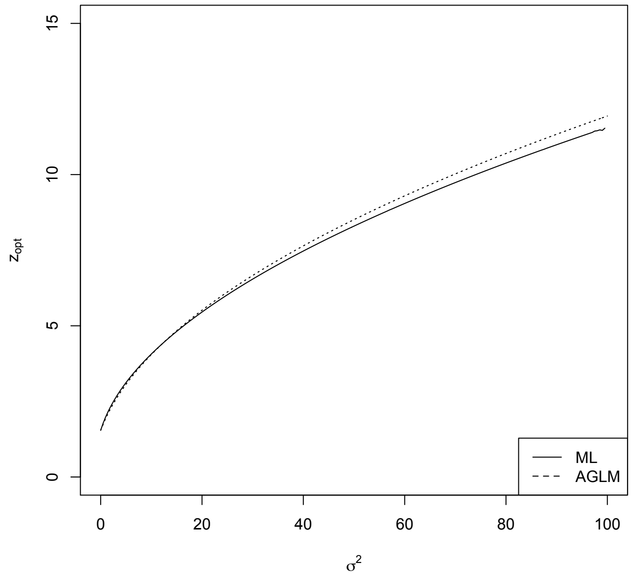

Figure 4.5: Locally optimal canonical designs under ML and AGLM, $\sigma^2 \leq 100$

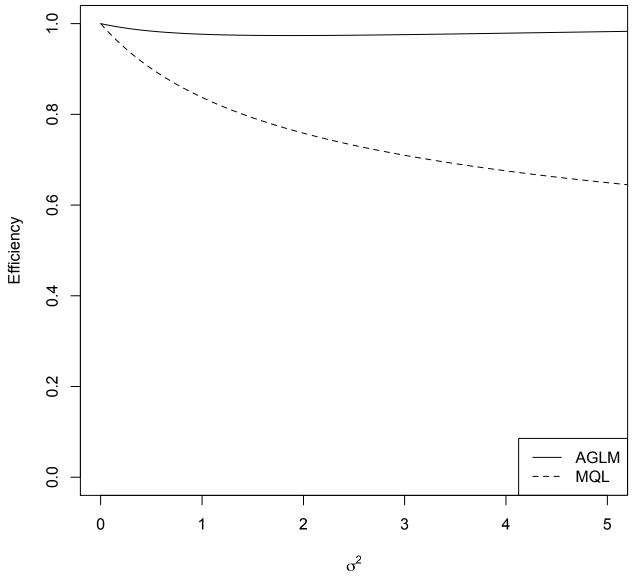

Figure 4.6: D-efficiency of MQL and AGLM locally optimal designs, as a function of $\sigma^2 \leq 5$

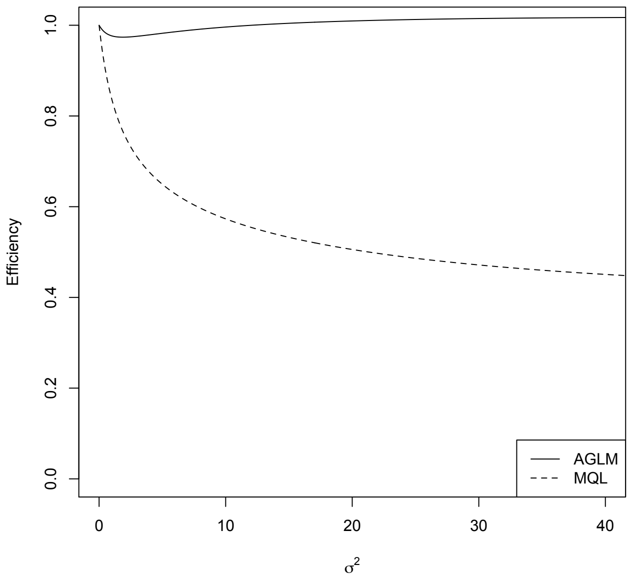

Figure 4.7: D-efficiency of MQL and AGLM locally optimal designs, as a function of $\sigma^2 \leq 40$

4.5 Robust designs

The optimal design depends on the unknown values of the model parameters. Before the experiment is performed there is uncertainty about the values of these parameters, which complicates the choice of a good experimental design. If we use a locally optimal design we must choose a best initial guess for the parameters, however if this guess is poor then the resulting design may be highly inefficient.

Ideally we would like an experimental design which is robust to different possible values of the parameters. In this section we calculate some more robust designs using two techniques. The first involves combining locally optimal designs from a number of plausible values of the parameters, and the second method is a Bayesian approach. The latter involves codifying uncertainty about the parameter values using a prior distribution, and then choosing the design which has the highest average efficiency with respect to those prior beliefs.

Maximum mean efficiency designs are not the standard Bayesian designs seen in the literature: typically one maximises the average log-determinant of the information matrix (e.g. Chaloner and Larntz, 1989). In Chapter 7 we give a detailed account of reasons for the use of maximum mean efficiency designs. Similarly, mixtures of locally optimal designs are not commonly employed, however they are extremely cheap to obtain once the locally optimal designs are available. Therefore they may serve as a crude benchmark.

Throughout Sections 4.5.1 and 4.5.2 we focus on the computation of robust optimal designs for the 1 point per block GLMM model (4.3). For illustration we use the MQL approximation to compute designs and efficiencies. In Section 4.5.3 we compare the results with optimal designs for the dose error model of Tang and Bacon-Shone (1992), which uses a probit link.

4.5.1 Mixtures of locally optimal designs

In this example we entertain, for illustration, four arbitrarily chosen possibilities $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_4$ for the parameter vector $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)$. These possibilities, together with the corresponding locally optimal designs $\xi_1, \ldots, \xi_4$ are listed in Table 4.1. The table of efficiencies for $\xi_1, \ldots, \xi_4$ under each of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_4$ is given in Table 4.2. From the latter, we see that the maximally robust locally optimal design (Melas, 2005) is ξ_1 which has a minimum efficiency of 7.8%, occurring under $\boldsymbol{\theta}_3$. It is clear that this worst-case performance of the design is in fact quite poor.

Recall that the function on the set of $p \times p$ real symmetric matrices defined by $M \mapsto |M|^{1/p}$ is concave (e.g. Firth and Hinde, 1997). Let us consider a weighted average of the locally optimal designs $\xi_1, \ldots, \xi_4$ defined by a vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_4)$ of positive weights such that $\sum_{i=1}^{4} \gamma_i = 1$. We define the weighted average of the designs $\xi_1, \ldots, \xi_4$,

$$\xi' = \sum_{i=1}^{4} \gamma_i \xi_i \,,$$

as follows: if x is a support point of at least one of the ξ_i then it is also a support point of ξ', and its probability mass in ξ' is the weighted average of its probability masses in $\xi_1, \ldots, \xi_4$, where the average is weighted by $\boldsymbol{\gamma}$.

Note that

$$\begin{aligned}
\mathrm{eff}(\xi';\boldsymbol{\theta}_j) &= \frac{|\sum_{i=1}^4 \gamma_i M_{\boldsymbol{\beta}}(\xi_i;\boldsymbol{\theta}_j)|^{1/2}}{|M_{\boldsymbol{\beta}}(\xi_j;\boldsymbol{\theta}_j)|^{1/2}} \\
&\geq \frac{\sum_{i=1}^4 \gamma_i |M_{\boldsymbol{\beta}}(\xi_i;\boldsymbol{\theta}_j)|^{1/2}}{|M_{\boldsymbol{\beta}}(\xi_j;\boldsymbol{\theta}_j)|^{1/2}} \\
&\geq \sum_{i=1}^4 \gamma_i \,\mathrm{eff}(\xi_i;\boldsymbol{\theta}_j) \\
&\geq (E^T\boldsymbol{\gamma})_j \,,
\end{aligned}$$

where the inequality follows due to concavity, and E is the 4×4 *efficiency matrix* whose ijth component is $\mathrm{eff}(\xi_i;\boldsymbol{\theta}_j)$. We rewrite this as the vector inequality

$$\mathrm{eff}(\xi';\Theta) \geq E^T\boldsymbol{\gamma}\,, \tag{4.28}$$

where Θ is the ordered set of plausible values of $\boldsymbol{\theta}$. The inequality (4.28) gives a lower bound on the efficiency under the different plausible values of $\boldsymbol{\theta}$. As the bound in (4.28) is linear in $\boldsymbol{\gamma}$, it can be manipulated rather easily. In particular, since E is invertible in this case, we can make the bound uniform across all the plausible $\boldsymbol{\theta}$ values by solving

$$E^T\boldsymbol{\gamma} = (1,\ldots,1)^T\,,$$

and rescaling so that $\sum_i \gamma_i = 1$. In this particular case, this yields

$$\boldsymbol{\gamma} = (0.03, 0.116, 0.415, 0.436)^T\,,$$

and the efficiency of ξ' is at least 45.1% in each case. Note that the weight on ξ_1 is small despite this being the maximally robust locally optimal design. This occurs because all of the designs have a reasonable performance under $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, whilst ξ_1 and ξ_2 perform badly under $\boldsymbol{\theta}_3$ and $\boldsymbol{\theta}_4$.

Clearly the design obtained in this way will always have a large number of support points, which may not be practical. However this design may provide a useful calibration point against which to measure the robustness of other designs. Moreover applying this approach may be useful when evaluating the information matrix for a given proposal design is costly, since in this case optimising a maximin or Bayesian objective function will be slow.

As an aside, the actual efficiency vector for the mixture design ξ' was $(0.779, 0.745, 0.559, 0.578)^T$, much larger than the lower bound.

	Parameter values				
	β_0	β_1	σ^2	Design points	
$\boldsymbol{\theta}_1$	0	1	1	-1.664	1.664
$\boldsymbol{\theta}_2$	2	1	0.5	-3.606	-0.394
$\boldsymbol{\theta}_3$	5	3	2	-2.255	-1.078
$\boldsymbol{\theta}_4$	-1	3	0.1	-0.185	0.852

Table 4.1: Plausible parameter values and locally optimal designs

	Parameter values			
Design	$\boldsymbol{\theta}_1$	$\boldsymbol{\theta}_2$	$\boldsymbol{\theta}_3$	$\boldsymbol{\theta}_4$
ξ_1	1.000	0.575	0.078	0.150
ξ_2	0.570	1.000	0.210	0.018
ξ_3	0.336	0.566	1.000	0.020
ξ_4	0.490	0.188	0.021	1.000

Table 4.2: Efficiencies of locally optimal designs

4.5.2 Bayesian designs

In this section we place a prior distribution on Θ, specifically $P(\boldsymbol{\theta} = \boldsymbol{\theta}_l) = \pi_l$, $l = 1, \ldots, 4$, with $\boldsymbol{\pi} = (0.2, 0.3, 0.25, 0.25)$. We then find the design ξ, with m support points, which maximises the mean efficiency,

$$\begin{aligned} E_{\boldsymbol{\theta}}\, \text{eff}(\xi; \boldsymbol{\theta}) &= \sum_{l=1}^{4} \pi_l\, \text{eff}(\xi; \boldsymbol{\theta}_l) \\ &= \sum_{l=1}^{4} \pi_l\, \frac{|M_{\boldsymbol{\beta}}(\xi; \boldsymbol{\theta}_l)|^{1/p}}{|M_{\boldsymbol{\beta}}(\xi_l; \boldsymbol{\theta}_l)|^{1/p}}\,. \end{aligned} \tag{4.29}$$

Note that the optimal determinant values in the denominator of (4.29) have already been computed in the previous subsection. With $k = 2$ the optimal design has a mean efficiency of 54%, and with 3–8 points the optimal mean efficiency is 68%. Compare this with the mixture design in the previous section, which has a mean efficiency of 66%. We are able to do better on average with just 3 (compared with 16) support points. The optimal 3-point Bayesian design is

$$\xi = \left\{ \begin{array}{ccc} -2.418 & -0.618 & 0.775 \\ 0.358 & 0.352 & 0.290 \end{array} \right\},$$

where the first row contains the design points and the second row gives the weights. The efficiency vector for ξ is

$$\text{eff}(\xi; \Theta) = (0.804, 0.782, 0.578, 0.561)^T\,.$$

In particular, the Bayesian design is more efficient than the mixture design ξ' under all of the plausible $\boldsymbol{\theta}$ values except $\boldsymbol{\theta}_4$, where it performs comparably.

In Section 7.5.2 we give a pseudo-Bayesian justification of the use of designs maximising (4.29), in the case where the analyst will perform a frequentist analysis whose value is to be measured in terms of size of confidence intervals produced.

It will often be desirable to use a more complex approximation to a continuous prior. In this case, numerical quadrature techniques such as that of Gotwalt et al. (2009) may be used.

4.5.3 Probit dose-error model

In Section 4.2.1, we defined Estimation Problems 1 & 2 for the random intercept and dose-error models respectively, and pointed out that the information matrices differ for these two problems. In this section, we calculate designs which are an approximation to the optimal designs for Estimation Problem 2, and compare these to the corresponding optimal designs for Estimation Problem 1.

Instead of the logit, we use a probit link function for the dose-error model since expressions for the information matrix in the latter case have been developed already by Tang and Bacon-Shone (1992). We also calculate robust designs for the probit dose-error model using the techniques of Sections 4.5.1 and 4.5.2.

Model and information matrix

The model considered by Tang and Bacon-Shone (1992) is parameterised in the following way

$$P(y=1|x,\epsilon) = \Phi(\beta(x+\epsilon-\gamma))\,, \tag{4.30}$$

where $\epsilon \sim N(0,\sigma_e^2)$ is the dose-error (with σ_e^2 known), Φ is the standard normal CDF, and β and γ are parameters to be estimated. This relates to

$$P(y=1|x,\epsilon) = \Phi(\beta_0+\beta_1(x+\epsilon))\,, \tag{4.31}$$

by the reparameterisation

$$\begin{aligned} \beta &= \beta_1 \\ \gamma &= -\beta_0/\beta_1\,. \end{aligned}$$

Note that (4.31) is simply (4.5) with a probit, rather than logit, link function. The nonlinear reparameterisation between (4.30) and (4.31) does not affect locally D-optimal designs, or local D-efficiencies, and so we use D-optimal designs for (4.30) as approximations to the D-optimal designs for (4.5).

Let ξ be an arbitrary approximate design with support log-doses $x_1,\ldots,x_k$ and weights $w_1,\ldots,w_k$. Let $\mathbf{\Gamma} = (\beta,\gamma)^T$. The information matrix for (4.30), with σ_ϵ^2 known, is (Tang and Bacon-Shone, 1992)

$$M_{\mathbf{\Gamma}}^{\mathrm{DE}}(\xi;\boldsymbol{\theta}) = \left(\begin{array}{cc} \sum_{i=1}^k \frac{w_i}{\Phi_i(1-\Phi_i)}\left(\frac{\partial\Phi_i}{\partial\gamma}\right)^2 & \sum_{i=1}^k \frac{w_i}{\Phi_i(1-\Phi_i)}\left(\frac{\partial\Phi_i}{\partial\gamma}\right)\left(\frac{\partial\Phi_i}{\partial\beta}\right) \\ \sum_{i=1}^k \frac{w_i}{\Phi_i(1-\Phi_i)}\left(\frac{\partial\Phi_i}{\partial\gamma}\right)\left(\frac{\partial\Phi_i}{\partial\beta}\right) & \sum_{i=1}^k \frac{w_i}{\Phi_i(1-\Phi_i)}\left(\frac{\partial\Phi_i}{\partial\beta}\right)^2 \end{array}\right),$$

where

$$\begin{aligned} \frac{\partial\Phi_i}{\partial\gamma} &= -\frac{\beta}{\sqrt{1+\beta^2\sigma_\epsilon^2}}\,\phi\left(\frac{\beta(x_i-\gamma)}{\sqrt{1+\beta^2\sigma_\epsilon^2}}\right) \\ \frac{\partial\Phi_i}{\partial\beta} &= \frac{x_i-\gamma}{(1+\beta^2\sigma_\epsilon^2)^{3/2}}\,\phi\left(\frac{\beta(x_i-\gamma)}{\sqrt{1+\beta^2\sigma_\epsilon^2}}\right) \\ \Phi_i &= \Phi\left(\frac{\beta(x_i-\gamma)}{\sqrt{1+\beta^2\sigma_\epsilon^2}}\right)\,. \end{aligned}$$

Plausible parameter values

We obtain the set of plausible parameter values for the probit dose-error model by applying the transformations

$$\begin{aligned} \beta &= \beta_1 \\ \gamma &= -\beta_0/\beta_1 \\ \sigma_\epsilon^2 &= \sigma^2/\beta_1^2 \,, \end{aligned}$$

to the plausible parameters for the logistic random intercept model in Table 4.1. Note that this is simply a reparameterisation from a random intercept model to a dose error model, and it does not attempt to use the best probit approximation of the logistic function.

Let us establish the notation $\boldsymbol{\kappa} = (\beta, \gamma, \sigma_\epsilon^2)$ to denote the vector containing the new parameters, and $\boldsymbol{\kappa}_1, \dots, \boldsymbol{\kappa}_4$ for the transformed parameter vectors corresponding to $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_4$. Then the locally optimal designs for the dose-error model are given in Table 4.3, together with the corresponding MQL designs from the logistic model. The designs for the dose-error model were calculated using the expressions of Tang and Bacon-Shone (1992). We note that the designs for the two models appear mainly to be quite similar. The efficiency of each optimal design from the random intercept model under the corresponding dose-error model is also shown. In most cases the random intercept designs were near-optimal under the dose-error model.

Efficiency matrix

Let us denote the locally optimal dose-error design at $\boldsymbol{\kappa}_i$ by ζ_i. The efficiency matrix, $E_{ij} = \text{eff}(\zeta_i|\boldsymbol{\kappa}_j)$, calculated using the dose-error model is

$$E = \begin{pmatrix} 1.000 & 0.299 & 0.002 & 0.000 \\ 0.424 & 1.000 & 0.273 & 0.000 \\ 0.359 & 0.670 & 1.000 & 0.000 \\ 0.383 & 0.106 & 0.004 & 1.000 \end{pmatrix},$$

from which we see that the maximally robust locally optimal design is ζ_4. This design may not be satisfactory, however, since under the worst case the efficiency is just 0.4.

Note that the efficiencies of $\zeta_1, \zeta_2, \zeta_3$ are very poor when the dose-error model parameters are equal to $\boldsymbol{\kappa}_4$. One might be tempted to suggest that this is because the differences between the random intercept and dose-error designs in Table 4.1 are greatest for $\boldsymbol{\kappa}_4$. However the observed low efficiency is due more to the choice of the dose-error model. To see this, we computed $\text{eff}_{\text{RI}}^{\text{ML}}(\zeta_1|\zeta_4, \boldsymbol{\theta}_4) = 0.18$, where $\text{eff}_{\text{RI}}^{\text{ML}}(\zeta_1|\zeta_4, \boldsymbol{\theta}_4)$ is the relative efficiency of designs ζ_1 and ζ_4 for estimating the random intercept, assuming parameter values $\boldsymbol{\theta}_4$. The ML approximation was used for this calculation. The point of this calculation is that the same pair of designs has a very different relative efficiency when compared under the dose-error and random intercept models.

Mixture design

The weight vector for the mixture design is $(0.356, 0.239, 0.140, 0.267)$, and the lower bound on the efficiencies is 42.8%.

Bayesian design

The 3-point Bayesian design which optimises the mean efficiency is

$$\zeta = \begin{pmatrix} -0.2966 & -2.4638 & 0.6715 \\ 0.3961 & 0.3744 & 0.2294 \end{pmatrix}.$$

Its efficiency vector is $\mathrm{eff}(\zeta|\boldsymbol{\kappa}) = (0.733, 0.776, 0.462, 0.532)^T$, with mean 0.628. To achieve a better mean efficiency of 63.6%, at least 5 points are required.

	Design points				
Parameters	Dose-error		Random intercept		Efficiency
$\boldsymbol{\theta}_1$ $(\boldsymbol{\kappa}_1)$	-1.61	1.61	-1.66	1.66	0.999
$\boldsymbol{\theta}_2$ $(\boldsymbol{\kappa}_2)$	-3.39	-0.61	-3.61	-0.39	0.977
$\boldsymbol{\theta}_3$ $(\boldsymbol{\kappa}_3)$	-2.32	-1.01	-2.26	-1.08	0.988
$\boldsymbol{\theta}_4$ $(\boldsymbol{\kappa}_4)$	-0.06	0.73	-0.19	0.85	0.913

Table 4.3: Locally optimal designs under dose-error model, with probit link, and random intercept model, with logistic link. The final column shows the D-efficiency of the random intercept design for estimating the dose-error model.

4.6 Robustness of estimation

In this section we investigate the impact of misspecifying σ^2 on the estimation of β_0 and β_1.

4.6.1 Recap of proposed estimation procedure

Recall that the proposed estimation procedure is to maximise the 'conditional' likelihood of β_0, β_1 which is obtained by plugging an assumed value, σ_g^2, of σ^2 into the actual (log)likelihood, in other words $\hat{\boldsymbol{\beta}}$ is the maximiser of

$$\ell(\beta_0, \beta_1; \sigma_g^2) = \sum_{i=1}^{n} \log \left\{ \int_{-\infty}^{\infty} h(\beta_0 + \beta_1 x_i + u_i)^{y_i} \left(1 - h(\beta_0 + \beta_1 x_i + u_i)\right)^{(1-y_i)} \phi_{\sigma_g^2}(u_i) du_i \right\}, \quad (4.32)$$

where $h(t) = 1/(1 + \exp(-t))$ is the logistic function, and ϕ_{σ^2} is the pdf of a $N(0, \sigma^2)$ random variable.

4.6.2 Computational details

We used a quasi-Newton method (BGFS, see e.g. Dennis and Schnabel, 1987, pp. 198–203) to directly maximise the numerically calculated (log)-likelihood. We supplied a routine for the calculation of the derivative of the log-likelihood (i.e. the score), to avoid difficulties experienced with the evaluation of finite-difference approximations to the derivative.

A notable feature is that precomputation of a lookup table for the marginal mean of the model speeds up the ML estimation by several orders of magnitude. In a specific example we encountered, the reduction in maximisation time was from 30s when integrals were evaluated just-in-time compared to 0.009s when precompuation was employed. Let us define the following

$$\begin{aligned} s_{\sigma^2}(\eta) &= (h * \phi_{\sigma^2})(\eta) \\ &= \int_{-\infty}^{\infty} h(\eta+u)\phi_{\sigma^2}(u)du \\ &= \int_{-\infty}^{\infty} h(\eta+\sigma u)\phi_1(u)du\,, \end{aligned} \tag{4.33}$$

where $*$ denotes the convolution operator. Note that if η is the value of the linear predictor for an individual, then the marginal expectation of their response is $s_{\sigma^2}(\eta)$. Then in fact the likelihood (4.32) can be evaluated in terms of s_{σ^2} as

$$\ell(\beta_0, \beta_1; \sigma_g^2) = \sum_{i=1}^{n} \log\left\{ y_i s_{\sigma_g^2}(\beta_0+\beta_1 x_i) + (1-y_i)\Big(1 - s_{\sigma_g^2}(\beta_0+\beta_1 x_i)\Big)\right\}. \tag{4.34}$$

There are two advantages to the use of (4.34) over (4.32). Firstly, the use of an addition, rather than exponentiation in the integrand is computationally faster. Secondly, if a table for $s_{\sigma_g^2}$ has been precomputed, changing the value of $\boldsymbol{\beta}$ incurs no extra integration with (4.34). We tabulate $s_{\sigma^2}(\eta)$, using the numerical quadrature provided by `integrate` in `R`, for η on a grid of spacing 0.1 over $[-10, 10]$. This was adequate for the values of σ_g^2 included in our study. Linear interpolation is used to approximate function values at values of η in between grid points.

The derivative, $\partial\ell/d\boldsymbol{\beta}$ required for the BFGS algorithm can be computed in the following way. Denoting the contribution to the likelihood from individual i by L_i we have that

$$\begin{aligned} \ell &= \sum_{i=1}^{n} \log L_i \\ L_i &= y_i s(\eta_i) + (1-y_i)\big(1 - s(\eta_i)\big) && (4.35) \\ \frac{\partial L_i}{\partial \eta_i} &= y_i s'(\eta_i) - (1-y_i) s'(\eta_i) && (4.36) \\ \frac{\partial \ell}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \frac{1}{L_i}\frac{\partial L_i}{\partial \eta_i} \begin{pmatrix} 1 \\ x_i \end{pmatrix}, && (4.37) \end{aligned}$$

where the final line holds by applying the chain rule twice. If we also precompute a lookup table for s', we can evaluate the score very quickly using equations (4.35)–(4.37). This is relatively simple, since

$$\begin{aligned} s'_{\sigma^2}(\eta) &= \int_{-\infty}^{\infty} h'(\eta+\sigma u)\phi_1(u)du \\ &= \int_{-\infty}^{\infty} \frac{e^{-(\eta+\sigma u)}}{(1+e^{-(\eta+\sigma u)})^2}\phi_1(u)du\,, \end{aligned}$$

which can be tabulated as easily as s_{σ^2}, using `integrate` with a grid of spacing 0.1 on $[-10, 10]$.

4.6.3 Results

We evaluated MSEs of $\hat{\beta}_0$ and $\hat{\beta}_1$ in the case where $(\beta_0, \beta_1) = (1, 1.5)$, and there are $n = 100$ individuals. Various combinations of true σ^2 and assumed σ_g^2 were considered.

In the first instance, a fixed design ξ_1 was assumed which assigns 50 individuals each to the doses $\log(x_{(1)}) = -5/3$ and $\log(x_{(2)}) = 1/3$. This design is close to D-optimal when $\sigma^2 = 0$, as it has linear predictors equal to ± 1.5. The calculated MSEs, based on 10,000 simulated datasets, for $\hat{\beta}_0$ and $\hat{\beta}_1$ are shown in Figures 4.8(a) and 4.9(a) respectively. Each curve corresponds to a different true value of σ^2, and the progression from left to right shows the change in MSE as the assumed value σ_g^2 changes. Note in particular that it is always optimal to take a value of σ_g^2 which is *less* than the true value of σ^2. However, the penalty incurred by taking $\sigma_g^2 = \sigma^2$, if the latter is known, is never very large. In the cases shown, the MSE is relatively robust for $\sigma_g^2 \geq \sigma^2 - 1.5$ or so. Choosing $\sigma_g^2 \leq \sigma^2$ is clearly better than choosing $\sigma_g^2 > \sigma^2$, thus it is better to use a guess for σ^2 which is an underestimate.

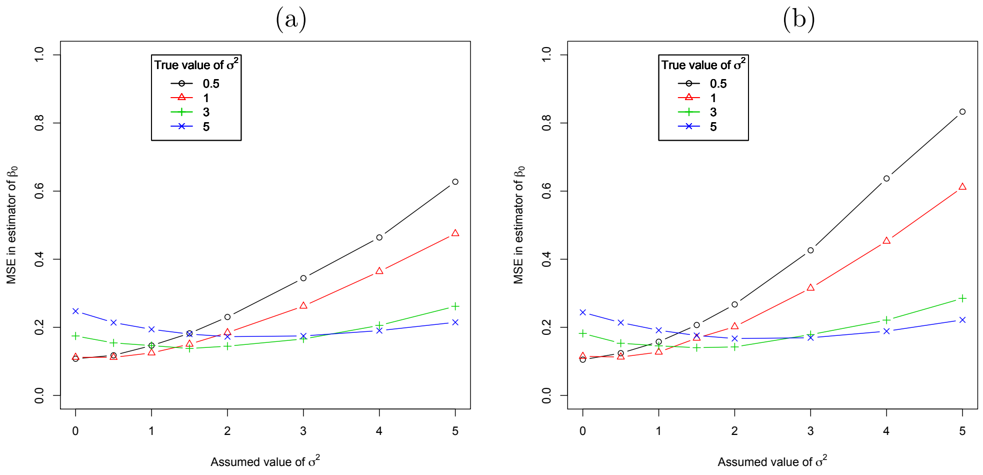

Figure 4.8: Mean squared error of $\hat{\beta}_0$, under (a) fixed design, and (b) near-optimal design strategies, for varying values of true σ^2 and assumed σ_g^2

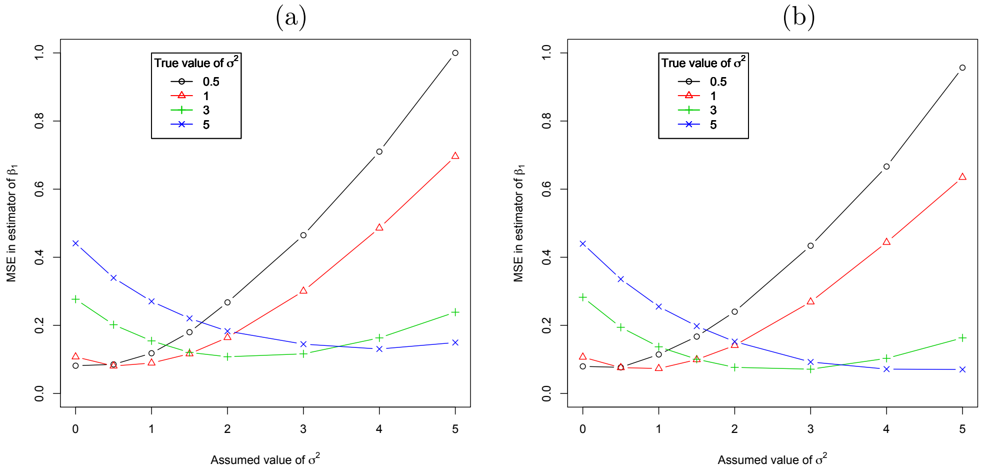

Figure 4.9: Mean squared error of $\hat{\beta}_1$, under (a) fixed design, and (b) near-optimal design strategies, for varying values of true σ^2 and assumed σ_g^2

A second study was conducted to reflect the fact that our assumed value of σ^2 may have an impact on the choice of design used. Specifically, given the assumed value σ_g^2, a probit approximation, $\xi^*(\beta_0, \beta_1, \sigma_g^2)$ to the locally optimal design (as in Section 4.4.4) was chosen. Note that the choice of this design is unrealistically efficient, since it requires knowledge of the true values of β_0 and β_1. Figures 4.8(b) and 4.9(b) show the MSEs for $\hat{\beta}_0$ and $\hat{\beta}_1$ under the combined design-estimation procedure. The difference made by using a locally optimal design at the guessed σ_g^2 is small. Figure 4.10 shows that the possible gains are biggest for $\hat{\beta}_1$ when σ^2 is large and σ_g^2 is close to the true value. When the guess is poor, the MSEs resulting from using the optimal design are worse. Thus, the optimal design strategy is slightly less robust to misspecification of σ_g^2 than using a fixed design. However, the factor which has the greatest bearing on the MSE is the choice of σ_g^2.

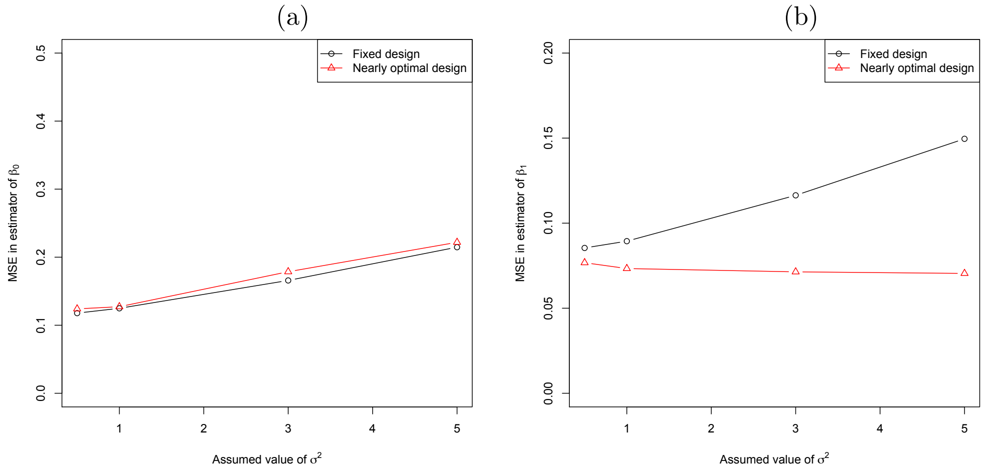

Figure 4.10: Mean squared error of (a) $\hat{\beta}_0$ and (b) $\hat{\beta}_1$, under fixed and near-optimal design strategies, for varying values of σ^2 and assuming we always guess correctly, $\sigma^2 = \sigma_g^2$.

4.7 Discussion

In this chapter we have examined the connection between the dose error model and the random intercept model with one point per block. Whilst the information matrix for these two models differs between Estimation Problems 1 and 2, we have found in Section 4.5.3 that often the locally optimal designs are similar under the two models.

By focussing on the logistic, rather than probit, model we have moved to a case where the relevant integrals are analytically intractable. However, the use of numerical integration in this problem is feasible due to the small number of variables and outcomes involved. This is in stark contrast to the case of the general GLMM of Chapter 2 where the approximations were required.

Despite the fact that it is not necessary to use MQL and PQL in this example, applying these approximations has given us insight into their relative performance. Clearly one would be much better using the ML designs or those resulting from the probit approximation for this problem. However, in keeping with our results in Chapter 3 we found that MQL is much better than PQL at locating the correct treatments for the optimal design. Once again, the PQL design had worse design points than the GLM design (which ignores the random effects) for $\sigma^2 > 0$.

We have been able to check that the optimal canonical design is of the assumed form (4.15): numerically in the case of ML, and analytically for MQL and PQL. An analytical proof for ML seems infeasible due to the presence of intractable integrals in the weight function.

The moderately robust designs formed from mixtures of locally optimal designs were cheap to obtain and serve as a benchmark for more complicated methods such as Bayesian designs. In practice their use will become difficult as uncertainty in the parameters increases; since one must consider a greater number of parameter values, the number of support points of the design will increase rapidly. Moreover with a continuous prior one must pick a small representative set of parameter values. The issue of the choice of this set would need further research before this method could be used more widely.

Maximum mean efficiency designs have a natural pseudo-Bayesian interpretation, and they avoid the assumption that the resulting data will be analysed using Bayesian methods. This is in contrast to approaches such as maximising the expected Kullback-Leibler divergence from the prior to the posterior, such as recommended by Chaloner and Verdinelli (1995). Moreover, maximum mean efficiency designs are fairly straightforward to obtain in simple models where there is a cheap way to obtain locally optimal designs. Such situations are becoming more common with the availability of analytical results for a wide range of two-parameter models (Konstantinou, Biedermann and Kimber, 2011), Poisson GLMs (Russell, Woods, Lewis and Eccleston, 2009), and other more general multifactor GLMs subject to restrictions on the ranges of the design variables (Yang et al., 2011). Maximum mean efficiency designs can also be helpful in the presence of singularities in the parameter space, for details see Chapter 7.

4.8 Appendix: Identifiability

In this section we show that for designs with a single dosing per individual, the model parameters are approximately unidentifiable.

We do this by noting that, given $\boldsymbol{\theta}' = (\beta_0', \beta_1', \sigma'^2)^T$, there is a (large) set, $\Theta(\boldsymbol{\theta}') \subseteq \mathbb{R}^2 \times (0, \infty)$, such that for $\tilde{\boldsymbol{\theta}} \in \Theta(\boldsymbol{\theta}')$,

$$P(Y = 1 \,|\, \tilde{\boldsymbol{\theta}}, x) \approx P(Y = 1 \,|\, \boldsymbol{\theta}', x) \quad \text{for all log-doses } x \in \mathbb{R}\,. \tag{4.38}$$

Specifically, such a set is given by

$$\Theta(\boldsymbol{\theta}') = \left\{ \tilde{\boldsymbol{\theta}} \;\middle|\; \tilde{\boldsymbol{\theta}} = (\gamma\beta_0', \gamma\beta_1', \tau^2)\,,\ \tau^2 > 0\,,\, \gamma = \sqrt{\frac{1 + c^2\tau^2}{1 + c^2\sigma'^2}} \right\},$$

where $c = 15\sqrt{3}/(16\pi)$. This follows from the approximation (4.4), since

$$\begin{aligned} P(Y = 1 \,|\, \tilde{\boldsymbol{\theta}}, x) &\approx \operatorname{expit}\left(\frac{\gamma\beta_0' + \gamma\beta_1' x}{\sqrt{1 + c^2\tau^2}} \right) \\ &= \operatorname{expit}\left(\frac{\beta_0' + \beta_1' x}{\sqrt{1 + c^2\sigma'^2}} \right) \\ &\approx P(Y = 1 \,|\, \boldsymbol{\theta}', x)\,. \end{aligned}$$

As a result of (4.38) it is hard to distinguish, with reasonable sample sizes, between putative parameter values in $\Theta(\boldsymbol{\theta}')$. One can derive analytical bounds on the difference $|P(Y = 1 \,|\, \boldsymbol{\theta}', x) - P(Y = 1 \,|\, \tilde{\boldsymbol{\theta}}, x)|$ for all $\boldsymbol{\theta}$, τ^2 and x. However these bounds are usually substantial overestimates – the difference being much less, particularly when τ^2 is moderately close to σ'^2. When considering sample size, we look at smaller differences.

Sample size. Consider an experiment whose purpose it is to detect whether $\boldsymbol{\theta} = \boldsymbol{\theta}'$ or $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$ best fits the data, for some particular fixed $\tilde{\boldsymbol{\theta}} \in \Theta(\boldsymbol{\theta}')$. We assume that the experiment consists of a single log-dose level, x, applied to all individuals, and that this dose gives optimal power for testing hypothesis

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}' \text{ vs. } H_1 : \boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$$

when using the one-sided rejection region

$$C = \left\{ \hat{p} \geq p(x; \boldsymbol{\theta}') + 1.6\sqrt{\frac{p(x;\boldsymbol{\theta}')[1 - p(x;\boldsymbol{\theta}')]}{n}} \right\} .$$

Above, $\hat{p} = \sum_{i=1}^{n} Y_i/n$ is the proportion of deaths, and

$$p(x; \boldsymbol{\theta}) = \int_{\mathbb{R}} h(\beta_0 + \beta_1 x + u)\phi_{\sigma^2}(u)\, du$$

is the marginal probability of death with these parameters and this dose. Note that we assume $p(x;\boldsymbol{\theta}') < p(x;\tilde{\boldsymbol{\theta}})$. Reversing this inequality and going through the argument changing signs leaves the final answer unchanged.

Under the assumption that n is large, the size of the above test is approximately 5%. Under H_1 the following event has approximately 95% probability,

$$I = \left\{ \hat{p} \geq p(x; \tilde{\boldsymbol{\theta}}) - 1.6\sqrt{\frac{p(x;\tilde{\boldsymbol{\theta}})[1 - p(x;\tilde{\boldsymbol{\theta}})]}{n}} \right\} ,$$

and $I \cap C = \emptyset$ if and only if

$$|p(x;\tilde{\boldsymbol{\theta}}) - p(x;\boldsymbol{\theta}')| \geq 1.6 \left[\sqrt{\frac{p(x;\tilde{\boldsymbol{\theta}})[1 - p(x;\tilde{\boldsymbol{\theta}})]}{n}} + \sqrt{\frac{p(x;\boldsymbol{\theta}')[1 - p(x;\boldsymbol{\theta}')]}{n}} \right] .$$

Therefore the test has 95% power to reject H_1 if and only if

$$n \geq (1.6)^2 \left[\sqrt{\frac{p(x;\tilde{\boldsymbol{\theta}})(1 - p(x;\tilde{\boldsymbol{\theta}}))}{n}} + \sqrt{\frac{p(x;\boldsymbol{\theta}')(1 - p(x;\boldsymbol{\theta}'))}{n}} \right]^2 |p(x;\tilde{\boldsymbol{\theta}}) - p(x;\boldsymbol{\theta}')|^{-2} .$$

We denote the RHS of the above inequality by $n_{\min}$. Note that with s_{σ^2} defined as in (4.33), and letting $\eta' = \beta_0' + \beta_1' x$, we have that

$$\begin{aligned} p(x;\boldsymbol{\theta}') &= s_{\sigma'^2}(\eta') \\ p(x;\tilde{\boldsymbol{\theta}}) &= s_{\tau^2}\left(\eta' \sqrt{\frac{1 + c^2\tau^2}{1 + c^2\sigma'^2}} \right) , \end{aligned}$$

so in fact $n_{\min}$ is determined completely by η', σ'^2 and τ^2. In other words one does not need to know the specific values of β_0', β_1', and x to compute $n_{\min}$.

As an example, consider the case where $\sigma^2 = 1$ and $\tau^2 = 2$. This yields $\gamma \approx 1.12$, which means that the values of the 'β' parameters in $\tilde{\boldsymbol{\theta}}$ are 12% inflated compared to those in $\boldsymbol{\theta}'$ (and so these parameter values are quite different). Figure 4.11 shows $n_{\min}$ as a function of η' for these values of σ'^2 and τ^2. The horizontal line shows $n = 4.5 \times 10^4$. If we choose x to minimise the required sample size then we still need more than 45,000 subjects to achieve 95% power. This figure will not be feasible in most applications. Note also that this dose level is unrealistically efficient, since to calculate the dose needed to achieve this value of η' we need to know the values of β_0, β_1.

Note as an aside that in the above example, we have $p(x; \boldsymbol{\theta}') = p(x; \tilde{\boldsymbol{\theta}})$ when $\eta' = 0$ and also when $\eta' \approx 2.74$. If an experiment uses these values of η', then the two parameter vectors will be completely unidentifiable from the data. Moreover if the value of η' used in the experiment is close to the above values, $n_{\min}$ can be arbitrarily large. This explains the presence of the vertical asymptote in Figure 4.11.

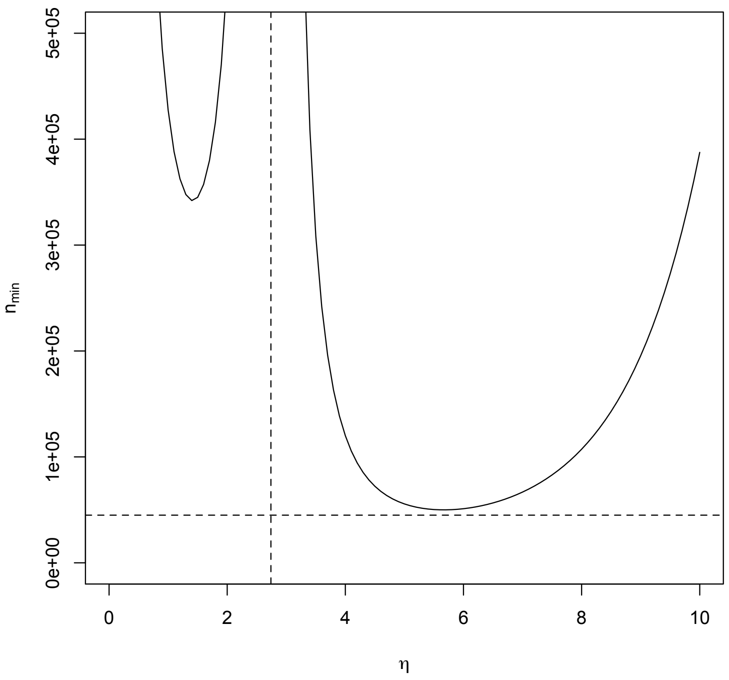

Figure 4.11: Sample size, $n_{\min}(\eta'; \sigma'^2, \tau^2)$, required for 95% power in test of θ' vs. $\tilde{\theta}$ as a function of η'.

4.9 Appendix: Information matrices for the two estimation problems

In this section we discuss why the information matrices differ for Estimation Problems 1 and 2 defined in Sections 4.2.1 and 4.2.2 respectively.

We begin by defining the following function, which will help us express the likelihood functions under models (4.3) and (4.5):

$$\begin{aligned} I(\eta, \sigma^2) &= \int_{-\infty}^{\infty} h(\eta + t)\, \phi_{\sigma^2}(t)\, dt \\ &= E_{t\sim N(0,\tau^2)} \left\{ h(\eta + t) \right\} \,, \end{aligned}$$

where ϕ_{σ^2} is the density of a $N(0, \tau^2)$ random variable, and h is the logistic function. The function $I(\eta, \sigma^2)$ is identical to $p_1(\eta, \sigma^2)$, defined in Section 4.4.1, but it is beneficial to have an unsubscripted version here.

Then the likelihood for the dose-error model, given a single observation at x_i which resulted in a 1, is given by

$$\begin{aligned} \ell^{\text{DE}}(\beta_0, \beta_1, \sigma_\epsilon^2 | 1) &= E_{\epsilon_i} \left[h(\eta_i + \beta_1 \epsilon_i) \right] \\ &= E_{\tilde{\epsilon}} \left[h(\eta_i + \tilde{\epsilon}_i) \right], \end{aligned}$$

where $\tilde{\epsilon}_i \sim N(0, \beta_1^2 \sigma_\epsilon^2)$ and $\eta_i = \beta_0 + \beta_1 x_i$. In terms of I, this is

$$\ell^{\text{DE}}(\beta_0, \beta_1, \sigma_\epsilon^2 | 1) = I(\beta_0 + \beta_1 x_i, \beta_1^2 \sigma_\epsilon^2) \,.$$

In contrast, the likelihood for the random intercept model (also given a 1 observed at dose x_i) is

$$\begin{aligned} \ell^{\text{RI}}(\beta_1, \beta_1, \sigma^2 | 1) &= E_{u_i} \left[h(\eta_i + u_i) \right] \\ &= I(\beta_0 + \beta_1 x_i, \sigma^2) \,. \end{aligned}$$

The key point is that when we come to evaluate the derivatives of the likelihoods with respect to β, we obtain different things for the two parameterisations. Using I_1 and I_2 to denote the partial derivative of I with respect to its first and second arguments respectively, we have by the chain rule that

$$\begin{aligned} \frac{\partial \ell^{\text{DE}}}{\partial \beta_1} =& I_1(\beta_0 + \beta_1 x_i,\, \beta_1^2 \sigma_\epsilon^2)\, x_i \\ & + I_2(\beta_0 + \beta_1 x_i,\, \beta_1^2 \sigma_\epsilon^2)\, 2\beta_1 \sigma_\epsilon^2 \,, \qquad (4.39) \\ \frac{\partial \ell^{\text{RI}}}{\partial \beta_1} =& I_1(\beta_0 + \beta_1 x_i,\, \sigma^2)\, x_i \,. \qquad (4.40) \end{aligned}$$

Using the fact that $\beta_1^2 \sigma_\epsilon^2 = \sigma^2$ makes the first term of (4.39) equal to (4.40). However there is no way of making the second term in (4.39) vanish. Moreover the second term does not vanish when these partial derivatives are substituted into (4.19) to form the information matrix.

4.10 Appendix: Proof of optimality of MQL design using symmetry arguments

Recall that the canonical locally optimal design problem is to maximise $\varphi(\mathbf{z}, \boldsymbol{\lambda}|\sigma^2)$, defined in (4.14), with respect to ξ_z. In this appendix, we give a proof that the optimal MQL design for the canonical problem is of the symmetric form

$$\xi^* = \begin{pmatrix} -z & z \\ 1/2 & 1/2 \end{pmatrix},$$

using symmetrisation argument, and a Lemma by Yang et al. (2011). A rather longer proof of this result, using the General Equivalence Theorem, is outlined in Section 4.11. The lemma is a slightly more general version of the result proved in Yang and Stufken (2009), and we quote it here, using slightly different notation:

Lemma 4.1 (Yang et al., 2011)**.** *Suppose that $F_1(c)$ and $F_3(c)$ are continuous functions on $(A, B]$ satisfying*

$$F_1'(c) < 0$$
$$\left(\frac{F_2'(c)}{F_1'(c)}\right)' > 0\,,$$

for $c \in (A, B]$. Then for any k given points $\{(c_i, w_i) : i = 1, \ldots, k\}$, where $c_i \in (A, B]$, $w_i \geq 0$, and $\sum_{i=0}^{k} w_i = 1$, there exists a point $\tilde{c}$ such that

$$\sum_{i=1}^{k} w_i F_1(c_i) = F_1(\tilde{c})\,,$$
$$\sum_{i=1}^{k} w_i F_2(c_i) < F_2(\tilde{c})\,.$$

Proof. The case $k = 2$ is equivalent to Proposition A.2 of Yang and Stufken (2009). Extend by induction, as done by Yang et al. (2011) for specific F_1 and F_2. □

Next we state some results on the Loewener ordering, which is fundamental in theoretical treatments of the notion of the information provided a design, see Pukelsheim (1987).

Definition 4.1 (Loewener ordering)**.** *Let A, B be symmetric $p \times p$ matrices. We say that $A \geq B$ in the Loewener ordering if the difference, $A - B$ is non-negative definite.*

Lemma 4.2 (Determinant respects Loewener ordering)**.** *If A, B are symmetric $p \times p$ matrices and $A \geq B$ in the Loewener ordering, then $|A| \geq |B|$.*

Proof. Let $\mathfrak{D}$ denote the function which maps a given symmetric $p \times p$ matrix M to $|M|^{1/p}$. By Firth and Hinde (1997), $\mathfrak{D}$ is concave. Thus

$$\mathfrak{D}(A) = 2\,\mathfrak{D}(\frac{1}{2}A)\,,$$

$$
\begin{aligned}
&= 2\,\mathfrak{D}\left(\frac{1}{2}B + \frac{1}{2}(A-B)\right) \\
&\geq 2 \times \left\{\frac{1}{2}\,\mathfrak{D}(B) + \frac{1}{2}\,\mathfrak{D}(A-B)\right\} \\
&\geq \mathfrak{D}(B)
\end{aligned}
$$

where the third line follows by concavity and the fourth line follows since, as $A-B$ is non-negative definite, $|A-B| \geq 0$. Taking pth powers gives the desired result. □

The final two lemmata before the main proof discuss symmetry properties, in the sense of Pukelsheim (1987), for the canonical problem. Let us denote an arbitrary design by

$$
\xi = \left(\begin{array}{ccc} z_1 & \dots & z_k \\ \lambda_1 & \dots & \lambda_k \end{array}\right),
$$

where the λ_i are weights and the z_i are support points, $i = 1, \dots, k$.

Lemma 4.3. *Let ξ be a design for the canonical problem, and $\mathcal{G} : z \mapsto -z\,, z \in \mathbb{R}$. Then $\mathcal{G}$ acts on the support points of ξ in the natural way, namely*

$$
\mathcal{G}(\xi) = \left\{\begin{array}{ccc} \mathcal{G}(z_1) & \dots & \mathcal{G}(z_n) \\ \lambda_1 & \dots & \lambda_n \end{array}\right\},
$$

and the determinant of the information matrix is unchanged by the application of $\mathcal{G}$. In other words $|M(\mathcal{G}(\xi))| = |M(\xi)|$. In the language of Pukelsheim (1987), the determinant, which is an information functional, is $\mathcal{G}$-invariant.

Proof. By (4.11), the information matrix of the transformed design is

$$
M(\mathcal{G}(\xi)) = \sum_{i=1}^{k} \lambda_i\, W(-z_i, \sigma^2) \left(\begin{array}{c} 1 \\ -z_i \end{array}\right) \left(\begin{array}{cc} 1 & -z_i \end{array}\right).
$$

As $W(-z, \sigma^2) = W(z, \sigma^2)$ for all the approximations, with some simple matrix algebra this can be re-expressed as

$$
\begin{aligned}
M(\mathcal{G}(\xi)) &= \sum_{i=1}^{k} \lambda_i\, W(z_i, \sigma^2) \left(\begin{array}{cc} 1 & 0 \\ 0 & -1 \end{array}\right) \left(\begin{array}{c} 1 \\ z_i \end{array}\right) \left(\begin{array}{cc} 1 & z_i \end{array}\right) \left(\begin{array}{cc} 1 & 0 \\ 0 & -1 \end{array}\right). \\
&= \left(\begin{array}{cc} 1 & 0 \\ 0 & -1 \end{array}\right) \left\{\sum_{i=1}^{k} \lambda_i\, W(z_i, \sigma^2) \left(\begin{array}{c} 1 \\ z_i \end{array}\right) \left(\begin{array}{cc} 1 & z_i \end{array}\right)\right\} \left(\begin{array}{cc} 1 & 0 \\ 0 & -1 \end{array}\right).
\end{aligned}
$$

Therefore, by the multiplicative properties of determinants,

$$
\begin{aligned}
|\mathcal{G}(M(\xi))| &= \left|\left(\begin{array}{cc} 1 & 0 \\ 0 & -1 \end{array}\right)\right| |M(\xi)| \left|\left(\begin{array}{cc} 1 & 0 \\ 0 & -1 \end{array}\right)\right| \\
&= |M(\xi)|\,.
\end{aligned}
$$

so $|M|$ is $\mathcal{G}$-invariant, as claimed. □

Lemma 4.4 (Designs are improved by symmetrisation)**.** *Define the symmetrised design by*

$$\begin{aligned}\xi_1 &= \left(\begin{array}{ccccc} z_1 & -z_1 & \dots & z_k & -z_k \\ \lambda_1/2 & \lambda_1/2 & \dots & \lambda_k/2 & \lambda_k/2 \end{array}\right) \\ &= (1/2)\xi + (1/2)\mathcal{G}(\xi)\,.\end{aligned}$$

Then the symmetrised design has a greater value of the objective function under D-optimality, in other words $|M(\xi_1)| \geq |M(\xi)|$.

Proof. We have that

$$\begin{aligned}\left|M\left(\frac{1}{2}\xi + \frac{1}{2}\mathcal{G}(\xi)\right)\right|^{1/p} &= \left|\frac{1}{2}M(\xi) + \frac{1}{2}\mathcal{G}(M(\xi))\right|^{1/p} \\ &\geq \frac{1}{2}|M(\xi)|^{1/p} + \frac{1}{2}|\mathcal{G}(M(\xi))|^{1/p} \\ &= |M(\xi)|^{1/p}\,,\end{aligned}$$

where the first line follows from additivity of the information matrix, the second from concavity of $|\cdot|^{1/p}$, and the third from Lemma 4.3. □

We now proceed to the proof of optimality of the symmetric design. As stated in Lemma 4.4, the design ξ can be improved upon by using the symmetrised design ξ_1. The information matrix of the symmetrised design is

$$M(\xi_1) = \left(\begin{array}{cc} \sum \lambda_i W(|z_i|, \sigma^2) & 0 \\ 0 & \sum \lambda_i |z_i|^2 W(|z_i|, \sigma^2) \end{array}\right).$$

It can be shown by relatively simple calculus that the functions $F_1(c) = W(c, \sigma^2)$ and $F_2(c) = c^2 W(c, \sigma^2)$ satisfy the conditions of Lemma 1 on $(0, \infty)$, where W is the MQL weight function defined in Section 4.4.2. Thus, taking $w_i = \lambda_i$ and $c_i = |z_i|$ in Lemma 1, there exists some $\tilde{c}$ such that

$$\sum_{i=1}^{k} \lambda_i W(|z_i|, \sigma^2) = W(\tilde{c}, \sigma^2) \tag{4.41}$$

$$\sum_{i=1}^{k} \lambda_i z_i^2 W(|z_i|, \sigma^2) < \tilde{c}^2 W(\tilde{c}, \sigma^2)\,. \tag{4.42}$$

Define a new design by

$$\tilde{\xi} = \left(\begin{array}{cc} \tilde{c} & -\tilde{c} \\ 1/2 & 1/2 \end{array}\right).$$

Then the information matrix for $\tilde{\xi}$ is

$$M(\tilde{\xi}) = \left(\begin{array}{cc} W(\tilde{c}, \sigma^2) & 0 \\ 0 & \tilde{c}^2 W(\tilde{c}, \sigma^2) \end{array}\right).$$

By (4.41) and (4.42), $M(\tilde{\xi}) \geq M(\xi_1)$ in the Loewner ordering. By Lemma 4.2, so too $|M(\tilde{\xi})| \geq |M(\xi_1)|$. As the symmetrised design, ξ_1, is better than ξ in the objective function sense, so is $\tilde{\xi}$. Hence there is an optimal design which consists of two symmetric points.

4.11 Appendix: Proof of optimality of MQL design using General Equivalence Theorem

In Section 4.4.2 we derived the optimal MQL design for given β and σ^2 using a canonical form. In the derivation, we assumed that the optimal design ξ was of the restricted form (4.15). Here we justify this assumption by using the General Equivalence Theorem to prove that the optimal symmetric design is in fact globally optimal.

4.11.1 General Equivalence Theorem

Recall that the locally D-optimal MQL canonical design,

$$\xi_z = \left\{ \begin{array}{ccc} z_1 & \dots & z_k \\ \lambda_1 & \dots & \lambda_k \end{array} \right\},$$

is that which maximises

$$\varphi(\xi_z; \sigma^2) = \varphi(\mathbf{z}, \boldsymbol{\lambda}; \sigma^2) = |M_{\boldsymbol{\beta}}(\xi_z; \boldsymbol{\theta}_c)|\,,$$

where $\boldsymbol{\theta}_c = (0, 1, \sigma^2)^T$. In the above, $M_{\boldsymbol{\beta}}$ is the MQL approximation to the information matrix. Let us define the derivative of $\log\varphi$ at ξ in the direction of a point (transformed log-dose) $z \in \mathcal{X}$ by the following:

$$d(z, \xi_z) = \lim_{\epsilon \to 0}(\epsilon^{-1}\left[\log\varphi\left\{(1-\epsilon)\xi_z + \epsilon\delta_z\right\} - \log\varphi(\xi_z)\right])\,,$$

where δ_z denotes the design which places unit mass at the point z. Then the General Equivalence Theorem (for example Chaloner and Larntz, 1989) states that

$$\xi^* \text{ maximises } \varphi(\xi_z) \text{ if and only if } \sup_{z \in \mathcal{X}} d(z, \xi^*) = 0\,.$$

Moreover, in this case, the derivative will attain the supremum at the support points of the design, in other words

$$d(z, \xi^*) = 0\,,\ \forall z \in \text{Support}(\xi^*)\,.$$

Note that the necessary property that the design region is compact does not hold here, since $\mathbb{R}$ is unbounded. However this problem can be circumvented in a way analogous to the one-factor logistic design problem without random effects (Chaloner and Larntz, 1989), namely by observing that the optimal design points are finite and restricting to an interval containing these points.

Note also that the General Equivalence Theorem has not been proved explicitly for MQL information matrices. However, the necessary property, that the information matrix is additive over independent experiments, clearly does hold.

4.11.2 Outline of proof

Let us use the following notation for the optimal symmetric design,

$$\xi = \left\{ \begin{array}{cc} -z_* & z_* \\ 0.5 & 0.5 \end{array} \right\}, \tag{4.43}$$

where z_* maximises $zW(z, \sigma^2)$, and W is the MQL weight function. In order to show that this symmetric design is in fact globally optimal for the canonical problem, we will show that $\sup d(z, \xi)$ is equal to zero. This is done in several parts, using calculus together with the fact that $d(z, \xi)$ has derivatives of all orders. Namely we will show that:

1. The Fréchet derivative $d(z, \xi)$ has at most 3 turning points. We demonstrate this essentially by showing that its fourth derivative, $d^{(4)}(z, \xi)$, is positive everywhere.
2. The support points $\pm z_*$ are turning points of $d(z, \xi)$, as is 0. By the above, these are all the turning points of $d(z, \xi)$.
3. The support points of ξ are zeroes of $d(z, \xi)$.
4. The derivative at zero, $d(0, \xi)$, is at most zero.
5. For z such that $|z| > |z_*|$, $d(z, \xi) < 0$.

These conditions are sufficient to establish that $\sup_{z \in \mathbb{R}} d(z, \xi) = 0$, by the following argument. By points 1, 2 and 3, the value of $d(0, \xi)$ is in fact strictly negative, since if it were zero then by Rolle's theorem there must be additional turning points in $(-z_*, 0)$ and $(0, -z_*)$, thus contradicting 1. Now suppose there were a point z_1 in $(-z_*, z_*)$ such that $d(z_1, \xi) > 0$. We may assume without loss of generality that $z_1 > 0$. Then by continuity there would also be $z_2 \in (0, z_1)$ such that $d(z_2, \xi) = 0$. By Rolle's theorem there would then be an additional turning point in $(0, z_*)$, contradicting 1. Thus $d(z, \xi) < 0$ for all z except $\pm z_*$.

To begin establishing properties 1–5 we must first obtain an expression for the function $d(z, \xi)$. Using a point prior in the expressions of Chaloner and Larntz (1989), we obtain

$$d(z, \xi_z) = \operatorname{tr}\left\{M^{-1}(\xi_z, \boldsymbol{\theta}) m(z, \boldsymbol{\theta})\right\} - p\,, \tag{4.44}$$

where p is the number of parameters of interest, in this case 2, and

$$m(z, \boldsymbol{\theta}) = M(\delta_z, \boldsymbol{\theta})$$

is the information matrix of the design which places unit mass at z. Recall that the canonical problem is defined by $\beta = (0, 1)$, and in this setup we have that

$$\begin{aligned} m(z, \boldsymbol{\theta}) &= W(z, \sigma^2) \begin{pmatrix} 1 & z \\ z & z^2 \end{pmatrix} \\ M(\xi, \boldsymbol{\theta}) &= W(z_*, \sigma^2) \begin{pmatrix} 1 & 0 \\ 0 & z_*^2 \end{pmatrix}, \end{aligned}$$

where in the second line we have used the particular form for ξ given by (4.43).

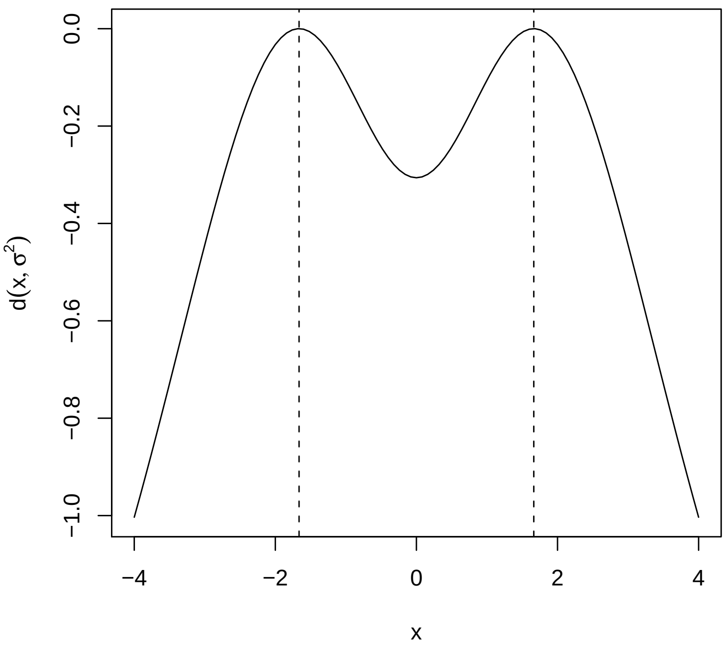

Figure 4.12: The derivative, $d(z,\xi)$, with $\sigma^2 = 1$. Vertical lines indicate $z = \pm z_*$

Substituting these matrices into (4.44), we obtain

$$d(z,\xi) = \frac{W(z,\sigma^2)}{W(z_*,\sigma^2)} \left(1 + \frac{z^2}{z_*^2}\right) - 2\,, \tag{4.45}$$

wherein the reason for the complicated proof becomes clear: the function d depends on constants z_* and $W(z_*,\sigma^2)$ which are defined only implicitly as the solution of the differential equation $\frac{d}{dz}(zw) = 0$. Moreover, it is inadequate to simply plot d as a function of z, since we require a proof for all values of σ^2. However, such a plot for a particular σ^2 serves as an indication of the general shape of d, see Figure 4.12.

In Sections 4.11.3–4.11.6 we establish properties 1–5.

4.11.3 Number of turning points

In this section we show that $d(z,\xi)$ has at most 3 turning points. First of all, let us derive an identity involving the first derivative of the weight function W. Recall from Section 4.4.2 that

$$1/W = e^z + 2 + e^{-z} + \sigma^2\,. \tag{4.46}$$

Therefore, by the chain rule,

$$-W^{-2}\frac{dW}{dt} = e^z - e^{-z}\,.$$

Thus,

$$\frac{dW}{dt} = W^2(e^{-z} - e^z)\,. \tag{4.47}$$

Now note that the $d(z, \xi)$, i.e. (4.45), may be rewritten as

$$\begin{aligned} d(z, \xi) &= \frac{1}{z_* W(z_*, \sigma^2)} W(z, \sigma^2) \left\{ z_*^2 + z^2 \right\} - 2 \\ &\propto W(z, \sigma^2) \left\{ z_*^2 + z^2 \right\} + \text{constant}. \end{aligned}$$

Therefore the first derivative of d with respect to z satisfies

$$\begin{aligned} \frac{d}{dz} \{d(z, \xi)\} &\propto z_*^2 \frac{dW}{dt} + z^2 \frac{dW}{dt} + 2zW \\ &\propto -W^2 (z_*^2 + z^2)(e^z - e^{-z}) + 2zW \\ &\propto -W^2 \left\{ (z_*^2 + z^2)(e^z - e^{-z}) - 2zW^{-1} \right\}. \end{aligned} \tag{4.48}$$

Hence z is a turning point of $d(z, \xi)$ if and only if $F(z) = 0$, where

$$\begin{aligned} F(z) &= (z_*^2 + z^2)(e^z - e^{-z}) - 2zW^{-1} \\ &= (z_*^2 + z^2)(e^z - e^{-z}) - 2z(e^z + 2 + e^{-z} + \sigma^2). \end{aligned} \tag{4.49}$$

We show that the third derivative of $F(z)$, $F'''(z)$, is positive everywhere. This is sufficient to show that F has at most 3 distinct zeroes. (Indeed, suppose there were 4 distinct zeroes, then we could repeatedly apply Rolle's Theorem to show the existence of 3, 2, and 1 distinct zeroes of F', F'', F''' respectively. The last of these contradicts the positivity of F'''). By basic calculus,

$$\begin{aligned} F'(z) &= e^z(z^2 + z_*^2 - 2) + e^{-z}(z^2 + z_*^2 - 2) - 2(2 + \sigma^2) \\ &= G(z) + G(-z) - 2(2 + \sigma^2), \end{aligned} \tag{4.50}$$

where we define G as

$$G(z) = e^z(z^2 + z_*^2 - 2).$$

Differentiating (4.50) twice, we obtain

$$F'''(z) = G''(z) + G''(-z). \tag{4.51}$$

It is relatively simple to check that

$$G''(z) = e^z(z^2 + 4z + z_*^2). \tag{4.52}$$

Substituting (4.52) into (4.51) yields

$$\begin{aligned} F'''(z) &= (e^z + e^{-z})(z^2 + z_*^2) + 4z(e^z - e^{-z}) \\ &> 0 \text{ for all } z. \end{aligned}$$

Thus $F(z)$ has at most 3 distinct roots, and $d(z, \xi)$ has at most 3 turning points.

4.11.4 Identifying turning points

Recall that $\pm z_*$ are such that $z_* W(z_*, \sigma^2)$ is maximised, therefore

$$\frac{d}{dz}(zW)\bigg|_{z=z_*} = 0\,. \tag{4.53}$$

However, by the product rule, we have

$$\frac{d}{dz}(zW) = W + z\frac{dW}{dz}\,. \tag{4.54}$$

Evaluating (4.48) at $z = z_*$ and using (4.54) we obtain

$$\begin{aligned}\frac{d}{dz}\left\{d(z,\sigma^2)\right\} &\doteq 2z_*^2\,\frac{dW}{dt}\bigg|_{z=z_*} + 2z_* W(z_*,\sigma^2)\\ &\doteq 2z_*\,\frac{d}{dz}(zW)\bigg|_{z=z_*}\\ &\doteq 0\,,\end{aligned}$$

where the dotted equals sign denotes equality up to multiplication by the constant of proportionality. As zero times a constant is zero, we have that $\pm z_*$ are turning points of $d(z, \sigma^2)$. Note also from (4.47) that $dW/dt = 0$ at $z = 0$. Therefore, evaluation of (4.48) at $z = 0$ yields zero. Hence 0 is also a turning point of $d(z, \sigma^2)$.

4.11.5 Value of $d(0, \sigma^2)$

Note that at $z = 0$, the derivative is

$$d(0,\xi) = \frac{W(0,\sigma^2)}{W(z_*,\sigma^2)} - 2\,.$$

Therefore to show that $d(0, \xi) \leq 0$, we must bound the value of W at z_*. Using (4.46) we obtain $W(0, \xi) = (4 + \sigma^2)^{-1}$. Therefore in particular we must show that

$$1/W(z_*,\sigma^2) \leq 2(4+\sigma^2)\,.$$

To do this, we make use of the following

Theorem 4.1. *The maximiser, $z_* > 0$, of $zW(z, \sigma^2)$ is the unique positive root of*

$$H(z) = 2 + \sigma^2\,,$$

where

$$H(z) = z(e^z - e^{-z}) - (e^z + e^{-z})\,. \tag{4.55}$$

Proof. By (4.47) and (4.54), it follows that

$$\frac{d}{dz}(zW) = W + zW^2(e^{-z} - e^z)\,.$$

Since $d(zW)/dz = 0$ at $z = z_*$, dividing the above by W^2 gives

$$\begin{aligned}0 &= W(z_*, \sigma^2)^{-1} + z_*(e^{-z_*} - e^{z_*}) \\ &= e^{z_*} + 2 + e^{-z_*} + \sigma^2 + z_*(e^{-z_*} - e^{z_*})\,.\end{aligned}$$

Rearranging gives that $H(z_*) = 2 + \sigma^2$. Uniqueness follows by considering the derivative, $H'(z) = z(e^z + e^{-z})$, which is positive for $z > 0$. □

Theorem 4.2. *The function $H(z)$ is bounded below as follows:*

$$H(z) \geq 2\cosh(z) - 4\,.$$

Proof. We proceed by expansion of the exponential functions in the definition of H. Note the following identities:

$$\begin{aligned}e^z + e^{-z} &= 2\left(1 + \frac{z^2}{2!} + \frac{z^4}{4!} + \dots\right) \\ &= 2 + 2\sum_{j=2,4,\dots} \frac{z^j}{j!} \\ z(e^z - e^{-z}) &= 2z\left(z + \frac{z^3}{3!} + \frac{z^5}{5!} + \dots\right) \\ &= 2\sum_{j=2,4,\dots} \frac{z^j}{(j-1)!}\,.\end{aligned}$$

Substituting the above in (4.55) yields

$$\begin{aligned}H(z) &= -2 + 2\left(\sum_{j=2,4,\dots} \frac{z^j}{(j-1)!} - \sum_{j=2,4,\dots} \frac{z^j}{j!}\right) \\ &= -2 + 2\sum_{j=2,4,\dots} \frac{j-1}{j!} z^j \\ &\geq -2 + 2\sum_{j=2,4,\dots} \frac{1}{j!} z^j \\ &\geq -2 + (2\cosh(z) - 2)\,.\end{aligned}$$

□

Theorem 4.3. *The value of w at z_* satisfies*

$$1/W(z_*, \sigma^2) \leq 2(4 + \sigma^2)\,.$$

Proof. By Theorems 4.1 and 4.2 above, at $z = z_*$ we have that

$$2 + \sigma^2 = H(z_*) \geq 2\cosh(z_*) - 4\,,$$

and so $\cosh(z_*) \leq (6+\sigma^2)/2$. However, by (4.46) it is also true that

$$\begin{aligned} 1/W(z_*, \sigma^2) &= e^{z_*} + 2 + e^{-z_*} + \sigma^2 \\ &= 2\cosh(z_*) + 2 + \sigma^2 \\ &\leq (6+\sigma^2) + (2+\sigma^2)\,, \end{aligned}$$

as required. □

4.11.6 Other properties

It remains to be shown that properties 3 and 5 from Section 4.11.2 hold. These properties are relatively simple to establish.

To prove property 3, that the support points, $\pm z_*$, of the optimal design are zeroes of the derivative, substitute $z = \pm z_*$ in (4.45). To demonstrate that property 5 holds, consider (4.45) and note that, for $|z| > |z_*|$

$$\begin{aligned} d(z, \xi) &= \frac{W(z,\sigma^2)}{W(z_*,\sigma^2)}\left(1 + \frac{z^2}{z_*^2}\right) - 2 \\ &< \frac{W(z,\sigma^2)}{W(z_*,\sigma^2)}\left(\frac{2z^2}{z_*^2}\right) - 2\,. \end{aligned} \tag{4.56}$$

By the definition of z_*, we have that

$$z^2 W(z, \sigma^2) \leq z_*^2 W(z_*, \sigma^2)\,, \tag{4.57}$$

for all z. Combining (4.56) and (4.57) we see that $d(z, \sigma^2) < 0$ for all z such that $|z| > |z_*|$, as required.

4.12 Appendix: Proof using BDZ

In this section, we prove optimality of the simplified form, (4.15), of the design under the MQL and PQL approximations. We also give numerical evidence to support that the simplified designs are also optimal under ML.

Biedermann et al. (2006), henceforth referred to as BDZ, consider binary outcome dose-response models where the probability of event occurrence for a particular dose level is

$$\pi(x) = H(\gamma(x-\alpha))\,, \tag{4.58}$$

with γ and α real-valued parameters, and H some cumulative distribution function on $\mathbb{R}$.

The information matrix of an approximate design ξ with weights λ_i and support doses x_i, $i = 1, \ldots, k$, is then

$$M(\xi) = \sum_{i=1}^{k} \lambda_i W(\gamma(x_i - \alpha)) \begin{pmatrix} \gamma^2 & -\gamma(x_i-\alpha) \\ -\gamma(x_i-\alpha) & (x_i-\alpha)^2 \end{pmatrix},$$

where

$$W(z) = \frac{(H')^2}{H(1-H)}(z)\,.$$

There is a class of design optimality criteria, referred to as Φ_p-optimality, $p \in (-\infty, 1]$, which is related to optimal estimation of $K^T(\alpha,\gamma)^T$, where K is a 2×2 matrix. This class of criteria includes the commonly-encountered A-, D- and E- optimality criteria (when $p = -1, 0, \infty$ respectively).

In their Theorem 2, Biedermann et al. (2006) derive a sufficient condition (called condition I) on W for the Φ_p-optimal design to be supported on two doses. This condition is that, for all $c \in \mathbb{R}$, the equation $(1/W)''(z) = c$ has at most two distinct roots. They state moreover, defining z to be $\gamma(x-\alpha)$, that if W is an even function with $W(z) = W(-z)$ then the transformed support points of the optimal design z_1, z_2 will be symmetric about 0, i.e. $z_1 = -z_2$.

Without changing any substantive details of the argument, it is possible to obtain a similar result when the information matrix is of the related form

$$M(\xi) = \sum_{i=1}^{k} \lambda_i W(z_i) \begin{pmatrix} 1 & z_i \\ z_i & z_i^2 \end{pmatrix}, \tag{4.59}$$

namely that when W satisfies condition I, the D-optimal design will be supported on at most 2 distinct points. Moreover, symmetry of W implies symmetry of the support points of the design. For completeness, we give the proof at the end of this Appendix.

We now note that the MQL and PQL information matrices are indeed of the form (4.59), and that their corresponding weight functions satisfy condition I.

MQL:

$$\begin{aligned} W_{\mathrm{MQL}}^{-1}(z) &= e^z + 2 + e^{-z} + \sigma^2 \\ (1/W_{\mathrm{MQL}})'' &= 2\cosh(z)\,. \end{aligned}$$

Clearly $(1/W)'' = c$ has at most two roots for all $c \in \mathbb{R}$. Indeed, there are precisely two roots for $c > 2$, one for $c = 2$ and none for $c < 2$.

PQL:

$$\begin{aligned} W_{\mathrm{PQL}}^{-1}(z) &= 2 + 2e^{\sigma^2/2}\cosh(z) + \sigma^2 \\ (1/W_{\mathrm{PQL}})''(z) &= 2e^{\sigma^2/2}\cosh(z)\,. \end{aligned}$$

Similarly, in this case $(1/W)'' = c$ has at most two roots for all $c \in \mathbb{R}$. Indeed, there are precisely two roots for $c > 2e^{\sigma^2/2}$, one for $c = 2e^{\sigma^2/2}$and none for $c < 2e^{\sigma^2/2}$.

By the above, we can verify with certainty that the MQL and PQL canonical designs are indeed of the simplified form (4.15). For ML the weight function is much more complicated and an analytical proof that W satisfies condition I remains elusive. However we can check numerically by making use of the following result, which is proved by elementary calculus (product rule and so on).

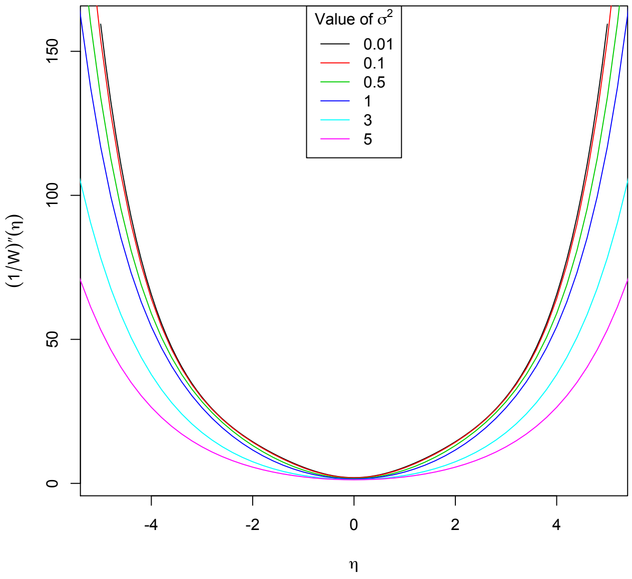

Figure 4.13: Numerical checking of condition I for the ML weight function, various σ^2.

Lemma 4.5. *Let* $W(z) = \frac{(H')^2}{H(1-H)}(z)$*. Then*

$$\frac{d^2}{dz^2}(1/W) = \frac{1}{(H')^4}\big\{ -6(H-1)H(H'')^2 \\ -2(H')^4 + 2(H-1)HH^{(3)}H' + 3(2H-1)(H')^2H''\big\} . \quad (4.60)$$

If $H(z) = s_{\sigma^2}(z)$ is defined as in (4.33), then we can evaluate s_{σ^2} and its derivatives using numerical integration, using

$$s^{(n)}_{\sigma^2}(z) = \int_{-\infty}^{\infty} h^{(n)}(z+u)\phi_{\sigma^2}(u)du\,,$$

with h the logistic function.

Figure 4.13 shows the numerically evaluated $(1/W)''$ from ML using various σ^2. In each case the condition seems to be satisfied, as $(1/W)''$ resembles a cosh type function.

Proof of result similar to BDZ, Theorem 2. We operate in less generality than Biedermann et al. (2006), however the proof is otherwise identical to that in their paper.

As we focus on D-optimal designs we can make use of the usual form of the General Equivalence theorem. We work again with the canonical problem, $\boldsymbol{\theta} = \boldsymbol{\theta}_c = (0, 1, \sigma^2)^T$. The directional derivative of $\log \varphi(\xi_z) = \log |M(\xi_z; \boldsymbol{\theta}_c)|$ in the direction of an arbitrary log-dose $z \in \mathcal{X}$ is

$$d(z, \xi_z) = \mathrm{tr}\{M^{-1}(\xi_z; \boldsymbol{\theta}_c)m(z; \boldsymbol{\theta}_c)\} - p\,,$$

where $m(z; \boldsymbol{\theta}_c)$ is the information matrix of a design supported only on z. Note the expression for $d(z, \xi_z)$ agrees with that in the previous Section.

Suppose that ξ is a D-optimal design, then $d(z, \xi_z) \leq 0$ for all $z \in \mathcal{X}$ with equality at the support points of ξ. Note that

$$m(z; \boldsymbol{\theta}_c) = W(z) \begin{pmatrix} 1 & z \\ z & z^2 \end{pmatrix},$$

and so

$$d(z, \xi_z) = W(z)\{a_0 + a_1 z + a_2 z^2\} - p \leq 0,$$

with equality at the support points. Equivalently,

$$q(z) \leq r(z), \tag{4.61}$$

with equality at the support points, where $q(z) = a_0 + a_1 z + a_2 z^2$, and $r(z) = p/W(z)$.

Suppose now that the support of ξ contains 3 distinct points, z_i, $i = 1, 2, 3$ with $q(z_i) = r(z_i)$. By the mean value theorem, there exist points z_i', $i = 1, 3$ with $z_1 < z_1' < z_2 < z_3' < z_3$ such that $q'(z_i') = r'(z_i')$, $i = 1, 3$. Moreover, by (4.61) we must have that $q'(z_i) = r'(z_i)$, $i = 1, 2, 3$. Thus we have 5 distinct points where q' and r' are equal. Applying the mean value theorem again (to q' and r'), we obtain 4 points where q'' and r'' are equal. As q'' is constant, this contradicts the assumption in condition I, namely that for all real c, $(1/W)'' = c$ has only 2 roots.

Thus there can be at most 2 distinct support points of any ξ which is D-optimal.

We now show that if W is symmetric, the design is symmetric in that it is supported on $\pm z$, for some $z \in \mathbb{R}$. Note that

$$M(\xi) = \begin{pmatrix} \sum_{i=1}^k \lambda_i W(z_i) & \sum_{i=1}^k \lambda_i z_i W(z_i) \\ \sum_{i=1}^k \lambda_i z_i W(z_i) & \sum_{i=1}^k \lambda_i z_i^2 W(z_i) \end{pmatrix},$$

and denoting by ξ^- the reflection of ξ at the origin, we have that

$$M(\xi^-) = \begin{pmatrix} \sum_{i=1}^k \lambda_i W(z_i) & -\sum_{i=1}^k \lambda_i z_i W(z_i) \\ -\sum_{i=1}^k \lambda_i z_i W(z_i) & \sum_{i=1}^k \lambda_i z_i^2 W(z_i) \end{pmatrix},$$

therefore $|M(\xi)| = |M(\xi^-)|$.

By concavity of φ, we have that $\xi_{\text{sym}} = (1/2)\xi + (1/2)\xi^-$ satisfies $\varphi(\xi_{\text{sym}}) \geq (1/2)\varphi(\xi) + (1/2)\varphi(\xi^-)$. By the preceding paragraph, the RHS is the optimal value of φ and so ξ_{sym} must also be D-optimal. Therefore ξ_{sym} must have 2 support points. By definition, ξ_{sym} is clearly symmetric, assigning equal weight to z and $-z$. This proves the symmetry result. □

Chapter 5

Multiple dosing designs

In order to be able to use the designs of Chapter 4, we need a prior estimate of σ^2, the individual random effect variance. We demonstrated in Section 4.6 that estimation of $\boldsymbol{\beta}$ is fairly robust to some misspecification of σ^2. However, if our a priori uncertainty about σ^2 is too large then the design strategy of using a single dosing per individual, together with the corresponding analysis, will be inadequate. As a consequence our design will need to be capable of estimating σ^2.

Another consideration is that σ^2 may not simply be a nuisance parameter, but it may be of interest in itself. For instance, we may be interested in understanding the selection effects described in Section 4.2.1. In this case, the ability to produce an estimate of σ^2 from the data is essential to the scientific question at hand.

In either of these two situations, we must be prepared to consider designs which involve multiple dosing events per individual. In this section, we investigate the optimal dose levels to be included in such designs.

5.1 Maximum likelihood estimation

For the ith individual, data takes the form of a sequence of (t_i-1) doses survived, $d_{i,1}, \ldots, d_{i,(t_i-1)}$, together with a final dose d_{i,t_i} and a variable y_i which indicates whether the individual is killed on the final dose. Letting h denote the expit, i.e. logistic, function, the log-likelihood for the individual is

$$\begin{aligned}\ell_{\text{indiv}}(\beta_0, \beta_1, \sigma^2; t_i, \mathbf{d}_i, y_i) = \log \Bigg\{ \int_{-\infty}^{\infty} & h(\eta_{i,t_i} + u)^{y_i} \left\{1 - h(\eta_{i,t_i} + u)\right\}^{1-y_i} \\ & \cdot \prod_{j=1}^{t_i-1} \left\{1 - h(\eta_{i,j} + u)\right\} \phi_{\sigma^2}(u)\, du \Bigg\}, \end{aligned} \tag{5.1}$$

where $\eta_{i,j} = \beta_0 + \beta_1 \log d_{i,j}$ is the linear predictor for the jth dose.

The log-likelihood given the complete data from n independent individuals is

$$\ell(\beta_0, \beta_1, \sigma^2; \mathbf{t}, \mathbf{D}, \mathbf{y}) = \sum_{i=1}^{n} \ell_{\text{indiv}}(\beta_0, \beta_1, \sigma^2; t_i, \mathbf{d}_i, y_i),$$

where $\mathbf{t}$ and $\mathbf{y}$ are the vectors consisting of the t_i and y_i respectively, and $\mathbf{D}$ is an array containing the d_{ij} (note that the rows of $\mathbf{D}$ are not necessarily the same length, and so the array is 'ragged'). For 'test to destruction' designs, $y_i = 1$ for all i.

We were able to implement a computational maximum likelihood estimation procedure for the full model by using numerical quadrature to perform evaluations of (5.1), together with generic numerical optimisation algorithms. Specifically, the functions `integrate` and `optim` in the `R` statistical computing environment were used (R Development Core Team, 2012). The BFGS algorithm was used, with numerically calculated derivatives.

To see whether this ML estimation procedure is effective, we performed a simple simulation study. With true parameter values $(\beta_0, \beta_1, \sigma^2) = (0, 1, 0.1)$ we generated 1000 datasets each consisting of 100 individuals. Each individual was dosed up to a maximum of 10^5 times (effectively until destruction). All individuals were assigned to the alternating two-dose sequence $\mathbf{d} = (d_1, d_2, d_1, d_2, \ldots)$ with $d_1 = 0.223$, $d_2 = 4.482$. The resulting Monte Carlo estimates for the means of the parameter estimators were $\tilde{\mathbb{E}}(\hat{\beta}_0) = 0.055 \pm 0.03$, $\tilde{\mathbb{E}}(\hat{\beta}_1) = 1.019 \pm 0.01$, $\tilde{\mathbb{E}}(\hat{\sigma}^2) = 0.105 \pm 0.008$. The numbers after the ± give the half-width, $1.96s/\sqrt{N_{mc}}$, of the approximate 95% Monte Carlo confidence interval, with s^2 the (unbiased) sample estimate of the variance of the parameter and N_{mc} the Monte Carlo sample size. The estimated means are very close to the true parameter values, suggesting the procedure works reasonably well. The confidence interval for $\mathbb{E}(\hat{\beta}_0)$ does not include the true value of β_0, but this does not contradict theory as the ML estimator using a sample size of 100 individuals is only approximately unbiased.

5.2 Evaluation of the information matrix

In this section we give details of the computation of the information matrix for general designs, and designs falling into an obvious restricted class.

Let $\mathbf{d}$ be a fixed dose sequence, of length m, which is to be followed until the event occurs. Thus the maximum possible number of dosing events is m, at which point the event has either occured or not. The possible outcomes for any individual are the times of death i.e. $t = 1, 2, \ldots, m, m^*$, where the asterisk denotes censoring, as in survival analysis. Given values of the parameters $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)^T$ let us denote the probability of outcome t by $P = P(t; \mathbf{d}, \boldsymbol{\theta})$. The set of possible outcomes for t is discrete, and finite, so that the individual information matrix can be calculated using a complete enumeration approach similar to that in Chapter 2. The resulting expression is

$$M_{\text{indiv}}(\mathbf{d}, m, \boldsymbol{\theta}) = \sum_{t=1,\ldots,m,m^*} -P(t; \mathbf{d}, \boldsymbol{\theta}) \frac{\partial^2 \log P(t; \mathbf{d}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \tag{5.2}$$

$$= \sum_{t=1,\ldots,m,m^*} \frac{1}{P(t; \mathbf{d}, \boldsymbol{\theta})} \left(\frac{\partial P(t; \mathbf{d}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial P(t; \mathbf{d}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^T . \tag{5.3}$$

Letting the predictor values be $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_m)^T$ we have

$$P(t; \mathbf{d}, \boldsymbol{\theta}) = \int_{-\infty}^{\infty} h(\eta_t + u) \prod_{j=1}^{t-1} \{1 - h(\eta_j + u)\} \, \phi_{\sigma^2}(u) \, du \,, \quad 1 \leq t \leq m \,, \tag{5.4}$$

where ϕ_{σ^2} is the density function of a $N(0,\sigma^2)$ random variable, and

$$P(m^*;\mathbf{d},\boldsymbol{\theta}) = \int_{-\infty}^{\infty} \prod_{j=1}^{m} \{1 - h(\eta_j + u)\}\, \phi_{\sigma^2}(u)\, du\,. \tag{5.5}$$

Note therefore that in fact $P(t;\mathbf{d},\boldsymbol{\theta}) = P(t;\boldsymbol{\eta},\sigma^2)$ for $t = 1,\ldots,m,m^*$. We will also use the notation $P_t(\boldsymbol{\eta},\sigma^2) = P(t;\boldsymbol{\eta},\sigma^2)$ when dependence on $\boldsymbol{\eta}$ and σ^2 is of prime importance.

For a general design, it is possible to evaluate the form (5.2) by a combination of evaluation of (5.4) and (5.5) using numerical quadrature, and numerical differentiation of $\log P$. When deriving optimal designs, we will work within the restricted class of designs featuring just one dose level per individual. In other words, we focus on designs such that the dose sequence corresponding to each individual is constant i.e. $\mathbf{d}_i = (d_i, d_i, \ldots, d_i)^T$. Within this class, considerable structure can be introduced in the form (5.3), which allows much faster computation of the information matrix. This is useful when comparing many candidate designs, as is necessary in the numerical search for optimal designs.

5.2.1 Evaluation for constant-dose-sequence designs

The form of (5.3) can be simplified in the restricted design setting by writing the partial derivatives $\partial P/\partial\boldsymbol{\theta}$ as integrals. This is done by taking the derivative under the integral sign (Theorem 3.1, Section 3.9), and it results in the individual information matrix being expressed in terms of a number of 'elementary integrals' $I_{j,t}(\eta;\sigma^2)$, described below, which are functions of η and σ^2 only.

For each value of η on a grid, the integrals $I_{j,t}$ and P_t are evaluated by use of a quadrature scheme. The result is that effectively all the integrals necessary to evaluate design properties for a particular σ^2 have been precomputed. Then the information matrix for different designs and parameter values can be evaluated simply by interpolation-type operations which are orders of magnitude faster than numerical evaluation of integrals.

For interpolation, we use splines whose coefficients are precomputed along with the values of the integrals. Splines were chosen because it is helpful to have a smooth approximation to the information matrix when performing optimisation. Initially, linear interpolation was used, but in this case the derivative functions in the equivalence theorem behaved rather oddly, having cusps at the grid points.

We consider the partial derivatives $\partial P_t/\partial\boldsymbol{\theta}$ separately for censored and uncensored outcomes;

Uncensored outcomes
Recall:

$$P_t(\boldsymbol{\eta};\sigma^2) = \int_{-\infty}^{\infty} h(\eta_t + u) \prod_{j=1}^{t-1} \{1 - h(\eta_j + u)\}\, \phi_{\sigma^2}(u)\, du\,.$$

We can evaluate the derivative of P_t with respect to the fixed effects parameters in the following way

$$\frac{\partial P_t}{\partial\beta_0} = \sum_{l=1}^{t} \frac{\partial P_t}{\partial\eta_l}\frac{\partial\eta_l}{\partial\beta_0}\,, \tag{5.6}$$

and similarly for β_1, by applying the chain rule from multivariable calculus to the decomposition,

$$\begin{array}{ccccc} \mathbb{R}^2 & \to & \mathbb{R}^m & \to & [0,1] \\ \boldsymbol{\beta} & \mapsto & \boldsymbol{\eta} & \mapsto & P_t(\boldsymbol{\eta}, \sigma^2)\,, \end{array} \tag{5.7}$$

of the mapping $\boldsymbol{\beta} \mapsto P_t(\boldsymbol{\eta}, \sigma^2)$. However, for $t \geq 2$ and $1 \leq l < t$,

$$\frac{\partial P_t}{\partial \eta_l} = \int_{-\infty}^{\infty} h(\eta_t + u)\{-h'(\eta_l + u)\} \prod_{1 \leq j \neq l \leq t-1} \{1 - h(\eta_j + u)\}\, \phi_{\sigma^2}(u)\, du\,.$$

This follows from a direct application of Theorem 3.1, similar to that in Section 3.9. In the case of a constant-dose-sequence design we have that $\eta_l = \eta$, for all l, in other words $\boldsymbol{\eta} = \eta\mathbf{1}$ is a constant vector. The above partial derivative, evaluated at a constant $\boldsymbol{\eta}$ is

$$\begin{aligned} \left.\frac{\partial P_t}{\partial \eta_l}\right|_{\boldsymbol{\eta}=\eta\mathbf{1}} &= (-1) \times \int_{-\infty}^{\infty} h'(\eta + u)\, h(\eta + u)\, \{1 - h(\eta + u)\}^{t-2}\, \phi_{\sigma^2}(u)\, du\,, \qquad \text{equal for all } l\,, \\ &=: I_{1,t}(\eta)\,, \end{aligned} \tag{5.8}$$

with $|$ denoting evaluation of the partial derivative function at a particular $\boldsymbol{\eta}$.

For $l = t$ we have

$$\begin{aligned} \left.\frac{\partial P_t}{\partial \eta_t}\right|_{\boldsymbol{\eta}=\eta\mathbf{1}} &= \int_{-\infty}^{\infty} h'(\eta + u)\{1 - h(\eta + u)\}^{t-1}\, \phi_{\sigma^2}(u)\, du \\ &=: I_{2,t}(\eta)\,, \end{aligned} \tag{5.9}$$

again following a straightforward application of Theorem 3.1. Substituting (5.8) and (5.9) into (5.6), we find that in the case of a constant dose individual design,

$$\frac{\partial P_t}{\partial \boldsymbol{\beta}} = \Big[(t-1)I_{1,t}(\eta) + I_{2,t}(\eta)\Big] \begin{pmatrix} 1 \\ \log d \end{pmatrix}. \tag{5.10}$$

Considering derivatives with respect to σ^2 yields

$$\begin{aligned} \left.\frac{\partial P_t}{\partial \sigma^2}\right|_{\boldsymbol{\eta}=\eta\mathbf{1}} &= \int_{-\infty}^{\infty} h(\eta + u)\{1 - h(\eta + u)\}^{t-1}\, \frac{\partial \phi_{\sigma^2}(u)}{\partial \sigma^2}\, du \\ &= I_{3,t}(\eta)\,, \end{aligned} \tag{5.11}$$

from a slightly more delicate application of Theorem 3.1, for details see Section 5.7.

Censored outcome

In this case the probability P is

$$P_{m^*}(\boldsymbol{\eta}, \sigma^2) = \int_{-\infty}^{\infty} \prod_{j=1}^{m} \{1 - h(\eta_j + u)\}\, \phi_{\sigma^2}(u)\, du\,, \quad 1 \leq t \leq m\,.$$

The derivatives of P with respect to η_l are

$$\frac{\partial P_{m^*}}{\partial \eta_l} = (-1) \times \int_{-\infty}^{\infty} h'(\eta_l + u) \prod_{1 \leq j \neq l \leq m} \{1 - h(\eta_j + u)\}\, \phi_{\sigma^2}(u)\, du\,.$$

and so, for all $1 \leq l \leq m$,

$$\begin{aligned} \left.\frac{\partial P_{m^*}}{\partial \eta_l}\right|_{\boldsymbol{\eta}} &= (-1) \times \int_{-\infty}^{\infty} h'(\eta + u)\,\{1 - h(\eta + u)\}^{m-1}\,\phi_{\sigma^2}(u)\,du\,, \qquad \text{equal for all } l\,, \\ &= -I_{2,m}(\eta)\,. \end{aligned} \tag{5.12}$$

Substituting (5.12) into the chain rule (5.6), we have that

$$\frac{\partial P_{m^*}}{\partial \boldsymbol{\beta}} = \Big[-mI_{2,m}(\eta)\Big]\begin{pmatrix} 1 \\ \log d \end{pmatrix}. \tag{5.13}$$

Finally, considering derivatives of P_{m^*} with respect to σ^2 gives

$$\begin{aligned} \left.\frac{\partial P_{m^*}}{\partial \sigma^2}\right|_{\boldsymbol{\eta}=\eta\mathbf{1}} &= \int_{-\infty}^{\infty} \{1 - h(\eta + u)\}^m \frac{\partial \phi_{\sigma^2}(u)}{\partial \sigma^2}\,du \\ &= I_{4,m^*}(\eta)\,. \end{aligned} \tag{5.14}$$

In common with the other derivatives taken with respect to σ^2 we must be slightly more careful with this application of Theorem 3.1, see Section 5.7.

If we wish to generate lookup tables for fixed maximum number of trials m then we need to tabulate, on a grid of η values, the following functions which are defined by the above integrals:

$$I_{1,t} \text{ for } t = 2, \ldots m$$

$$\left.\begin{matrix} P_t \\ I_{2,t} \\ I_{3,t} \end{matrix}\right\} \text{for } t = 1, \ldots, m$$

$$I_{4,m^*}$$

If for example $m = 20$, there are 80 functions to be tabulated. With a grid of step length 0.1 on $[-10, 10]$, it takes around 10s to perform all the necessary precomputations including the calculation of spline coefficients.

5.3 Locally optimal designs

In this section we work to find locally D-optimal designs, which maximise $\log|M_{\boldsymbol{\theta}}(\xi; \boldsymbol{\theta})|$. We first define a standardisation to a canonical form. This enables optimal designs to be found independently of the values of β_0 and β_1.

Throughout the rest of the chapter, we work with the notion of an approximate design. Let us assume that there are at most m dosing events per individual. In general, an approximate design for the multiple dosing problem is defined by a discrete probability measure on the set of dose-sequences of length m. In other words, an arbitrary design can be written as

$$\xi = \left\{ \begin{matrix} \mathbf{d}_1 & \ldots & \mathbf{d}_n \\ w_1 & \ldots & w_n \end{matrix} \right\},$$

where $\mathbf{d}_i = (d_{i1}, \ldots, d_{im})^T$ with $d_{ij} > 0$ for $i = 1, \ldots, n$, $j = 1, \ldots, m$, and the weights w_i satisfy $w_i > 0$, $\sum_{i=1}^n w_i = 1$. The interpretation of ξ is that a proportion w_i of the available individuals will follow dose sequence $\mathbf{d}_i$ until either death or the sequence is completed. However, we only derive constant-dose-sequence designs in which $\mathbf{d}_i = (d_i, \ldots, d_i)^T$. In this case we use the shorthand

$$\xi = \left\{ \begin{array}{ccc} d_1 & \ldots & d_n \\ w_1 & \ldots & w_n \end{array} \right\},$$

and it is taken as understood that the doses are to be repeated up to m times.

The overall information matrix for the design ξ is obtained from a weighted sum of the individual information matrices,

$$M(\xi; \boldsymbol{\theta}) = \sum_{i=1}^{n} w_i M_{\text{indiv}}(\mathbf{d}_i, m, \boldsymbol{\theta}).$$

5.3.1 Standardisation

Similarly to the case in Chapter 4 with one observation per individual, it is possible to define a linear transformation which relates the optimal design for arbitrary values of β_0 and β_1 to the optimal design in the case where (β_0, β_1) are equal to their 'canonical' values, $(0, 1)$.

Once we have established this result, all that remains is to numerically evaluate optimal designs for the canonical $\boldsymbol{\beta}$ and varying values of σ^2, which we do in Section 5.3.2. The first step in the proof is to observe a further structural property of the individual information matrix for constant-dose-sequence designs. Essentially we wish to decompose the individual information matrix into a product of three terms such that (i) the outer terms only depend on $\boldsymbol{\beta}$, and (ii) the central term depends on the design and the parameters, but only through η.

We first develop some additional notation for useful combinations of the elementary integrals. For $t = 1, \ldots, m$, define

$$\begin{aligned} w_t(\eta) &= (t-1) I_{1,t}(\eta) + I_{2,t}(\eta) \\ v_t(\eta) &= I_{3,t}(\eta), \end{aligned}$$

and also set

$$\begin{aligned} w_{m^*}(\eta) &= -m I_{2,m}(\eta) \\ v_{m^*}(\eta) &= I_{4,m^*}(\eta). \end{aligned}$$

Furthermore let

$$B = \begin{pmatrix} 1 & 0 & 0 \\ \beta_0 & \beta_1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and let us define the standardised individual information matrix to be

$$\tilde{M}(\eta) = \sum_{t=1,\ldots,m,m^*} \frac{1}{P_t(\eta)} \mathbf{q}_t(\eta) \mathbf{q}_t^T(\eta),$$

where

$$\mathbf{q}_t = \begin{pmatrix} w_t(\eta) \\ \eta w_t(\eta) \\ v_t(\eta) \end{pmatrix} .$$

Note that $M_{\text{indiv}}(\mathbf{d}, m, \boldsymbol{\theta}) = \tilde{M}(\eta)$ in the case where $\boldsymbol{\beta} = (0,1)^T$, and $\mathbf{d} = (d, d, \dots, d)^T$.

Lemma 5.1. *For* $\mathbf{d}$ *any constant dose-sequence, and* $\boldsymbol{\theta}$ *an arbitrary parameter vector, the individual information matrix can be written as*

$$M_{indiv}(\mathbf{d}, m, \boldsymbol{\theta}) = B^{-1}\tilde{M}(\eta)B^{-T} .$$

Proof. From results on constant-dose-sequence designs in equations (5.10), (5.11), (5.13), and (5.14) together with the definitions above, we have that

$$\frac{\partial P_t(\eta)}{\partial \boldsymbol{\theta}} = \begin{pmatrix} w_t \\ w_t \log d \\ v_t \end{pmatrix} .$$

Therefore by straightforward matrix multiplication, and the fact that $\eta = \beta_0 + \beta_1 \log d$,

$$B\frac{\partial P_t}{\partial \boldsymbol{\theta}} = \mathbf{q}_t . \tag{5.15}$$

Recall also that

$$M_{\text{indiv}} = \sum_{t=1,\dots,m,m^*} \frac{1}{P_t} \left(\frac{\partial P_t}{\partial \boldsymbol{\theta}}\right) \left(\frac{\partial P_t}{\partial \boldsymbol{\theta}}\right)^T ,$$

Provided $\beta_1 \neq 0$ we can premultiply (5.15) by B^{-1} and substitute the result into the complete enumeration equation above. This gives

$$\begin{aligned} M_{\text{indiv}} &= \sum_t \frac{1}{P_t} B^{-1}\mathbf{q}_t\mathbf{q}_t^T B^{-T} \\ &= B^{-1}\left\{\sum_t \frac{1}{P_t}\mathbf{q}_t\mathbf{q}_t^T\right\} B^{-T} . \end{aligned}$$

The term in the middle is precisely the definition of $\tilde{M}$. The result follows immediately. □

Lemma 5.2. *The locally D-optimal design for arbitrary* $\boldsymbol{\beta}$ *can be obtained by a straightforward transformation of the design* (η_i, w_i), $i = 1, \dots, n$, $\sum_{i=1}^n w_i = 1$, *solving the canonical optimisation problem*

$$\textit{maximise } \Upsilon(\boldsymbol{\eta}, \boldsymbol{w}) = \log \left| \sum_{i=1}^n w_i \tilde{M}(\eta_i) \right| .$$

The required transformation is given by $\log d_i = \beta_1^{-1}(\eta_i - \beta_0)$.

Proof. By Lemma 5.1, we have for any design ξ, with support points d_i and weights w_i, $i = 1, \dots, n$, that

$$M(\xi; \boldsymbol{\beta}, \sigma^2) = B^{-1}\left\{\sum_{i=1}^n w_i\tilde{M}(\eta_i)\right\} B^{-T} ,$$

and so

$$|M| = |B|^{-2} \left| \sum_{i=1}^{n} w_i \tilde{M}(\eta_i) \right| .$$

As $|B|$ does not depend on the design, maximising $|M|$ is equivalent to maximising the standardised information matrix of the 'predictor design' given by (η_i, w_i). □

5.3.2 Numerical results for various σ^2

In this section we calculate locally optimal designs with canonical $\boldsymbol{\beta}$ for various σ^2. As explained in Section 5.3.1, the dependence on $\boldsymbol{\beta}$ can be overcome using a canonical transformation. Combining the results from these two sections, we therefore have available the locally optimal designs in a broad range of parameter scenarios.

The optimal designs are computed using a 'co-ordinate optimisation' approach, restricting the search to designs supported on two dose levels. The optimality of the resulting designs is then verified using the General Equivalence Theorem, confirming the adequacy of designs supported on two doses. For further details of the algorithm and the General Equivalence Theorem in this context, see Section 5.4.2. Note that it is indeed possible to estimate all three parameters with such a design, which would not be the case if all the parameters were fixed effects parameters. As one of the parameters is a variance component, this estimability does not conflict with the classical theory.

The support doses of the locally D-optimal designs for various σ^2 in the range $[0.5, 5]$ are shown on a log-scale in Figure 5.1. The pattern is that as σ^2 increases, the optimal doses move further apart. This tendency towards more extreme doses parallels the situation in designs with one dosing per individual. In that case the phenomenon arises due to the attenuation in the marginal effect of $\log x$ introduced by the individual variation. A similar attenuation factor clearly applies here for the first dosing event.

The proportion of individuals allocated to the low dose in the locally optimal design is shown in Figure 5.2 as σ^2 varies. This weight increases from 0.35 to 0.41 as σ^2 increases from 0.5 to 5.

Figure 5.3 shows the shape of the derivative function from the equivalence theorem for each of the locally optimal designs computed. These plots verify (up to numerical approximation errors) the optimality of the designs found, as the derivative is in each case never much bigger than zero, and has approximate zeroes at the support points of the design.

5.4 Bayesian designs

In this section we give some examples of the use of precomputed interpolation tables for constructing Bayesian designs for the multiple dosing problem.

5.4.1 Objective function evaluation

We find designs which maximise an approximation to the objective function of Chaloner & Larntz (1989), in other words

$$\psi(\xi) = E_{\boldsymbol{\theta}}\{\log |M(\xi; \boldsymbol{\theta})|\} , \tag{5.16}$$

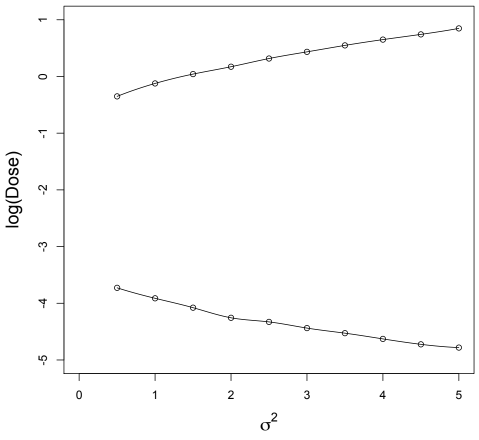

Figure 5.1: Support points of the locally optimal designs, with $(\beta_0, \beta_1)^T = (0, 1)^T$, as σ^2 varies in $[0.5, 5]$. Note the log-transformed vertical axis.

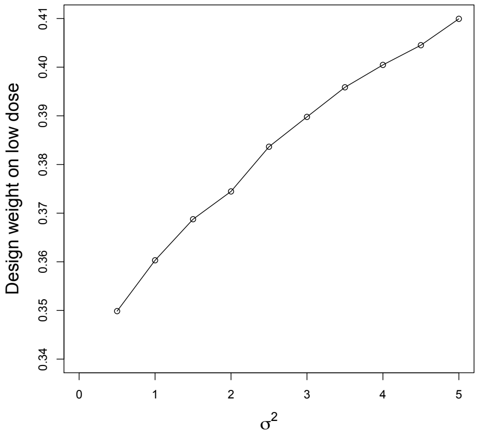

Figure 5.2: Weight of the lowest dose in the locally optimal designs with $(\beta_0, \beta_1)^T = (0, 1)^T$, as σ^2 varies in $[0.5, 5]$. Note that the vertical axis ranges from 0.34 to 0.41.

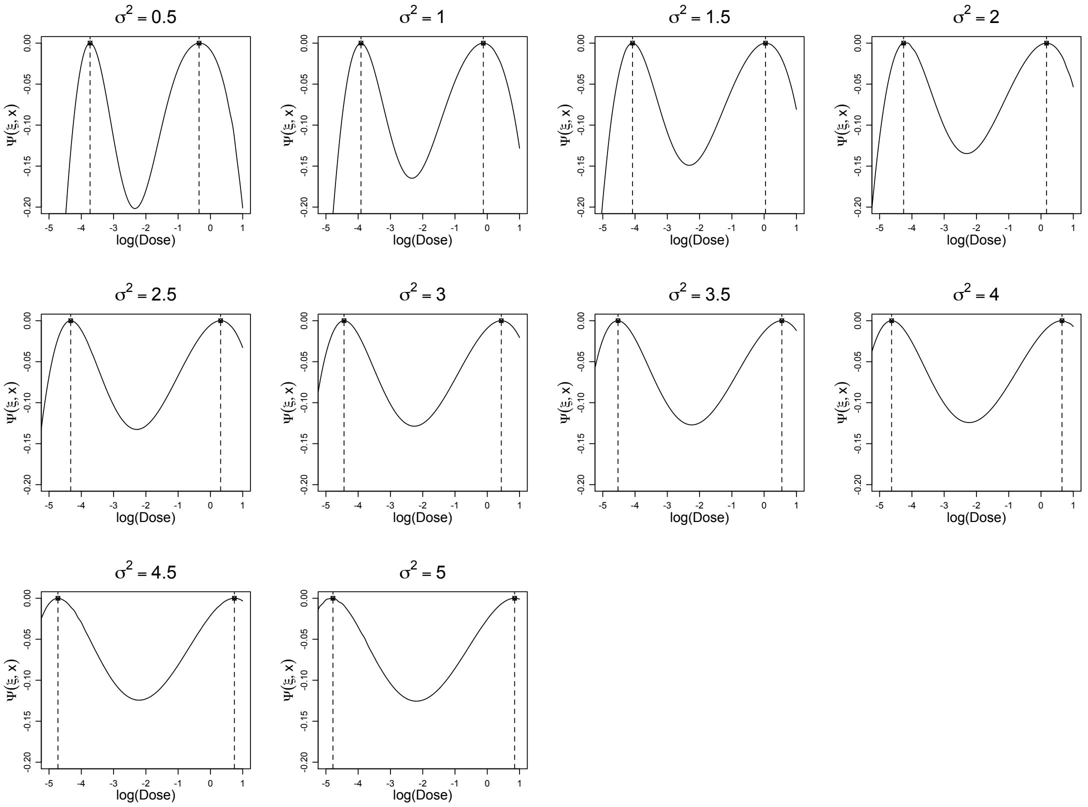

Figure 5.3: Plots of the GET derivative function $\Psi(\xi_i^*, d)$, (5.19), of the locally optimal designs for $\sigma_i^2 = 0.5, 1, 1.5, \ldots, 5$. The panels correspond to the different values of σ^2. Dotted vertical lines indicate the location of the support points of the design.

where the expectation is with respect to the prior distribution on $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)^T$. Usually the prior distribution has a density function, $f(\boldsymbol{\theta})$. Note however that if the prior distribution assigns unit mass to a particular parameter vector, then we recover the objective function for local D-optimality.

As a computational surrogate for (5.16), we employ the discretised version

$$\psi_d(\xi) = \sum_{s=1}^{N_a} \pi_s \log |M(\xi; \boldsymbol{\theta}_s)| \,, \tag{5.17}$$

where for $s = 1, \ldots, N_a$, $\boldsymbol{\theta}_s = (\beta_{0s}, \beta_{1s}, \sigma_s^2)^T$ and π_s are the integration abscissae and weights respectively. The weights satisfy $\sum_s^{N_a} \pi_s = 1$. Details of the formation of the abscissae and weights are given in the examples.

When computing (5.17) for many candidate designs, it is advantageous to tabulate the elementary integrals $I_{j,t}(\eta; \sigma^2)$, $j = 1, \ldots, 4$, $t = 1, \ldots, m$, defined in Section 5.2.1. This precomputation is performed on a grid of η values for each value of σ^2 appearing in the abscissae, in other words for $\sigma^2 = \sigma_s^2$, $s = 1, \ldots, N_a$.

5.4.2 Optimisation

To derive the designs in this section we use a combination of a 'co-ordinate optimisation' algorithm, and verification of optimality using the General Equivalence Theorem. Recall that a constant-dose-sequence design ξ can be written as

$$\xi = \left\{ \begin{array}{ccc} d_1 & \ldots & d_n \\ w_1 & \ldots & w_n \end{array} \right\} ,$$

and that ξ assigns a proportion w_i of the individuals to dose d_i and repeatedly applies this dose up to a maximum of $m = 20$ times.

The co-ordinate optimisation algorithm proceeds by iteratively adjusting each of $d_1, \ldots, d_n$ and then $w_1, \ldots, w_n$. When a new value of a dose is proposed, all other design parameters are held constant. When a new value of the weight w_i is proposed, the ratios between w_j, $j \neq i$ are held constant, and these weights are adjusted to preserve the constraint $\sum_{k=1}^n w_k = 1$. Thus on setting $w_i \leftarrow w_i'$, one sets the remaining weights as $w_j \leftarrow w_j(1 - w_i')/(1 - w_i)$. At each step, the new value of the design parameter is selected to maximise the value of (5.17). Thus the design search consists of a sequence of one-dimensional optimisation problems. Let us consider ϕ_d as a function of d_i ceteris paribus, writing $\varphi_i(d_i) = \psi_d(\xi)$. Then $\varphi_i(d_i)$ typically has several local maxima. We attempt to avoid choosing a local optimum which is not optimal over all d_i by the following method. First we compute $\varphi_i(v_j)$, $j = 1, \ldots, k$, on a grid, $v_1, \ldots, v_k$, of potential d_i spanning a wide range of values. We then note j such that $\varphi_i(v_j)$ is maximised, and concentrate a more refined optimisation on a neighbourhood of v_j.

We now discuss the use of the General Equivalence Theorem in this context. The derivative of ψ, at design ξ in the direction of an arbitrary alternative design ζ is defined to be

$$\Psi(\xi; \zeta) = \lim_{\alpha \to 0} \alpha^{-1} [\psi\{(1 - \alpha)\xi + \alpha\zeta\} - \psi(\xi)] \,,$$

and with ψ as in (5.16) this can be evaluated as

$$\Psi(\xi;\zeta) = E_{\boldsymbol{\theta}}\,\mathrm{tr}\{M^{-1}(\xi;\boldsymbol{\theta})M(\zeta;\boldsymbol{\theta})\} - p\,, \tag{5.18}$$

see for instance Firth & Hinde (1997) or Atkinson et al. (2007, Section 18.2). In the above, tr denotes the trace of a matrix, and p is the number of parameters contained in M, which is therefore of order $p \times p$. We can approximate (5.18) by

$$\Psi_d(\xi;\zeta) = \sum_{s=1}^{N_a} \pi_s\,\mathrm{tr}\{M^{-1}(\xi;\boldsymbol{\theta}_s)M(\zeta;\boldsymbol{\theta}_s)\} - p\,,$$

and this is in fact the derivative of the discretised objective function (5.17). This expression can be evaluated numerically given numerically evaluated information matrices $M(\xi;\boldsymbol{\theta}_s)$, $M(\zeta,\boldsymbol{\theta}_s)$.

We make some additional definitions before stating the Theorem. For $d > 0$, let us define $\delta(d)$ to be the design which assigns unit mass to the dose sequence $d\mathbf{1}$, which repeats dose d a total of m times. Moreover let us define the shorthand $\Psi(\xi;d) = \Psi(\xi;\delta(d))$. We are now ready for the result.

Theorem 5.1. *If an objective function ψ is concave, which indeed (5.16) and (5.17) are, then the following statements concerning the design ξ^* are equivalent:*

1. *The design ξ^* is optimal, in other words $\psi(\xi^*) = \sup_\xi \psi(\xi)$.*
2. *For all potential doses $d > 0$, we have $\Psi(\xi^*;d) \le 0$.*

Moreover, at the support doses of ξ^* it is the case that $\Psi(\xi;d) = 0$.

The Theorem can be used to check the optimality of a given design. If the proposed design turns out to be suboptimal, then evaluation of Ψ can suggest ways of improving the current ξ, for instance as in Atkinson (2008). Usually we should consider including in the support those doses d where $\Psi(\xi, d)$ is at its maximum. Once a good set of support doses for ξ has been found, the doses can be held fixed and the weights optimised (again using a co-ordinate type procedure as above).

Note that by choosing a prior distribution which assigns point mass to $\boldsymbol{\theta}$ we see that Theorem 5.1 applies also to local D-optimality. In this case, the derivative is

$$\Psi(\xi;d) = \mathrm{tr}\{M^{-1}(\xi;\boldsymbol{\theta})M(\delta(d);\boldsymbol{\theta})\} - p\,. \tag{5.19}$$

5.4.3 Example 1

Assume the following discrete prior distribution

$$\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)^T = \begin{cases} (0, 1, .1)^T \text{ with probability } 1/3 \\ (0, 2, .4)^T \text{ with probability } 1/3 \\ (1, .5, 1)^T \text{ with probability } 1/3 \end{cases},$$

and use the support points of the distribution as the integration abscissae, with equal weights. This rather small prior at least has the feature of multiple values of σ^2.

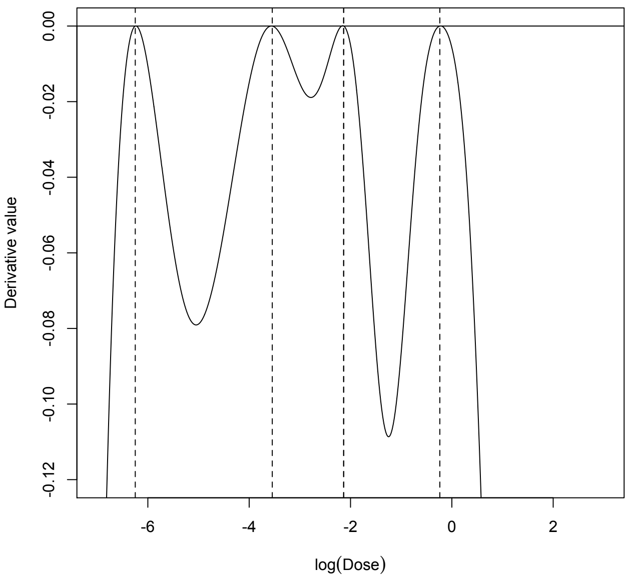

Figure 5.4: Derivative function $\Psi(\xi; d)$ for the optimal Bayesian design in Example 1 (Section 5.4.3). Vertical dotted lines indicate the location of the support points of the design. Note the log-transformed horizontal axis.

The optimal Bayesian design found was

$$\xi = \left\{ \begin{array}{cccc} 0.002 & 0.029 & 0.118 & 0.789 \\ 0.136 & 0.005 & 0.326 & 0.532 \end{array} \right\} .$$

A plot of the derivative from the General Equivalence Theorem is given in Figure 5.4. This confirms the optimality of the design: the function is non-positive and attains zero at the support points of the design (up to numerical approximation errors).

On its own, the co-ordinate algorithm struggled to identify the dose $d = 0.029$, due to the small optimal weight. However, omitting this dose still results in a design which is close to being optimal. Let ξ' denote the design which sets the weight of interest to zero, and multiplicatively rescales the remaining weights to sum to 1. Then $\psi(\xi) = -4.275826$ and $\psi(\xi') = -4.275897$.

5.4.4 Example 2

Here we assume the following independent prior distributions for the model parameters,

$$\begin{aligned} \beta_0 &\sim U(-0.5, 0.5) \\ \beta_1 &\sim U(0.6, 1.4) \\ \sigma^2 &\sim U(0, 3) \, . \end{aligned}$$

We compute the objective function using abscissae generated from a 30-point random Latin hypercube sample. The abscissae were evenly weighted.

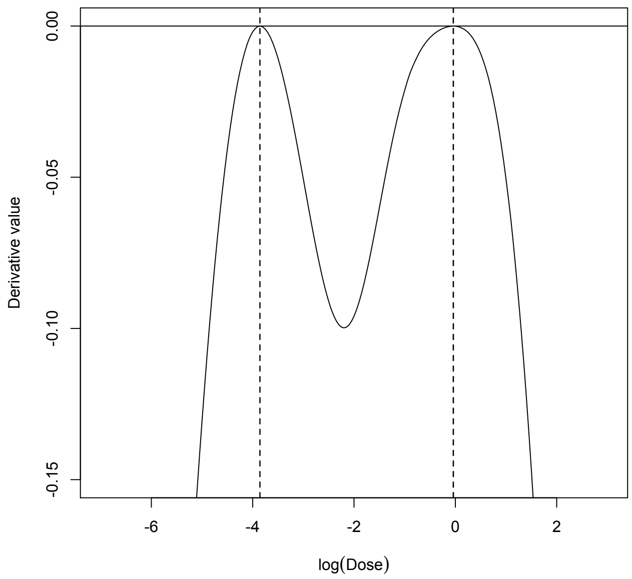

Figure 5.5: Derivative function $\Psi(\xi; d)$ for the optimal Bayesian design in Example 2 (Section 5.4.4). Vertical dotted lines indicate the locations of the support points of the design.

The optimal design supported on 4 doses was found to be

$$\xi = \left\{ \begin{array}{cccc} 0.021137 & 0.021134 & 0.957547 & 0.964165 \\ 0.161461 & 0.222379 & 0.354727 & 0.261433 \end{array} \right\} .$$

Its Equivalence Theorem derivative is shown in Figure 5.5. The derivative seems to satisfy the conditions for the design to be near-optimal. Clearly, the doses fall into two pairs of almost equal doses and it should be adequate to consolidate the doses within these pairs. Let ξ' denote the design obtained by rounding the doses to three decimal places, which consolidates the first two doses. Then $\psi(\xi) = -5.53642$ and $\psi(\xi') = -5.53644$, so there is little difference between ξ and ξ' in terms of performance.

This example was computationally much more involved. Due to the larger number of abscissae, evaluation of the objective function took of the order of a second.

5.5 Designs not tailored for σ^2

Let us define

$$M_{\boldsymbol{\beta}}(\xi; \boldsymbol{\theta}) = \sum_{i=1}^{n} w_i \, \mathbb{E}_{T_i} \left\{ \frac{-\partial^2 \log P(T_i; \mathbf{d}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right\} ,$$

where T_i is the (random) number of doses administered, before death, to an individual receiving dose sequence $\mathbf{d}_i$. Defined thus, $M_{\boldsymbol{\beta}}(\xi; \boldsymbol{\theta})$ is the part of the information matrix corresponding to β_0 and β_1. This reduced information matrix has a natural interpretation. If σ^2 is known, and we perform an experiment using design ξ with N individuals then

$$\mathrm{var}(\hat{\beta}) \approx (1/N) M_{\boldsymbol{\beta}}^{-1}(\xi; \boldsymbol{\theta})\,.$$

We use this to motivate a further optimality criterion. We say that ξ is *$D_{\boldsymbol{\beta}}$-optimal* if it maximises the value of $|M_{\boldsymbol{\beta}}(\xi; \boldsymbol{\theta})|$. A $D_{\boldsymbol{\beta}}$-optimal design optimises (asymptotically) point estimation of $\boldsymbol{\beta}$ when σ^2 is known.

In this section, we look at $D_{\boldsymbol{\beta}}$-optimal designs for the unit variation model, with the same restrictions as in Section 5.3. In Chapter 2, we developed approximations to the equivalent of $M_{\boldsymbol{\beta}}$ for GLMMs. We briefly consider the performance of these approximations for the unit variation model here.

It is relatively straightforward to use the work of the previous sections to obtain designs which optimise $|M_{\boldsymbol{\beta}}|$. One simply ignores the third row and column of the information matrices evaluated using complete enumeration. The same optimisation algorithms can be applied to yield the locally optimal design, in other words co-ordinate optimisation restricted to designs supported on two doses. In the parameter scenario $(\beta_0, \beta_1, \sigma^2)^T = (0, 1, 1)^T$ we found the optimal doses were 0.02 and 2.56, both evenly weighted.

For local $D_{\boldsymbol{\beta}}$-optimality, the objective function is $\psi(\xi) = \log|M_{\boldsymbol{\beta}}(\xi; \boldsymbol{\theta})|$, and the directional derivative from the General Equivalence Theorem is

$$\Psi(\xi; d) = \mathrm{tr}\{M_{\boldsymbol{\beta}}^{-1}(\xi; \boldsymbol{\theta}) M_{\boldsymbol{\beta}}(\delta(d); \boldsymbol{\theta})\} - 2\,.$$

This follows from (5.18) since the information matrix $M_{\boldsymbol{\beta}}$ is 2×2. Figure 5.6 shows the computationally evaluated derivative for the $D_{\boldsymbol{\beta}}$-optimal design found from a numerical search, which satisfies the required conditions. The maximal dose in the D-optimal design for this scenario is smaller than 1 (i.e. lower than the 'standard' dose), in contrast the $D_{\boldsymbol{\beta}}$-optimal design uses a dose which is larger than 1.

Note that the model is no longer a straightforward GLMM, due to the stopping rule applying at the individual level (i.e. that we can take no further observations on an individual once they have died). This makes the assumptions underlying the MQL approximation to $\mathrm{var}\,\hat{\boldsymbol{\beta}}$ more questionable. We do not return to first principles to derive a new approximation. Instead, we try to use the existing approximation to produce a design which might be applicable in this situation. Specifically, we derive an MQL-optimal wholeplot design for a 1-factor GLMM with $m = 20$ points per block, using the same assumed values of $\boldsymbol{\beta}, \sigma^2$ as above. We can use the same optimisation algorithms as for the complete enumeration designs, and this yields optimal doses 13.74, 0.07 which are again evenly weighted. The derivative function again confirms optimality of this design (Figure 5.7).

The D-efficiency of the MQL design relative to the $D_{\boldsymbol{\beta}}$-optimal complete enumeration design is 80.0%. To see why the MQL design is inefficient, note that the maximal dose used is around 14 times the 'standard' unit dose, which is much larger than in any of our previous designs. The likely reason such high doses were not present in the complete enumeration designs is that the latter designs acknowledge the stopping rule. At a high dose, individuals are very likely to die on the first attempt: we can extract more information by using a lower dose which allows the individual to survive longer and thereby provide more observations.

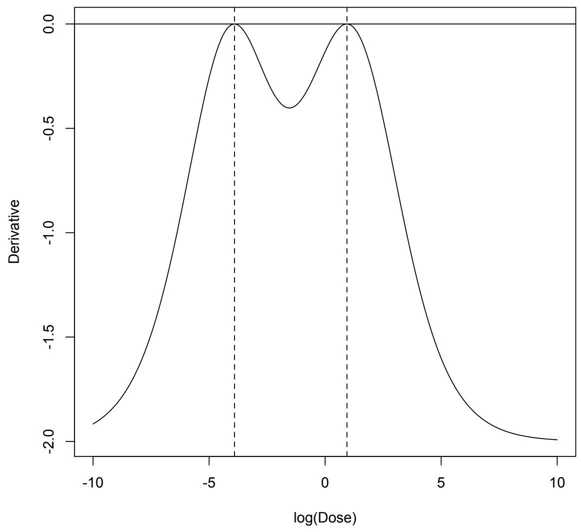

Figure 5.6: Derivative function for $D_{\boldsymbol{\beta}}$-optimal design, with information matrices evaluated using complete enumeration. Vertical dotted lines show the location of the support doses. Note doses are plotted on the log scale.

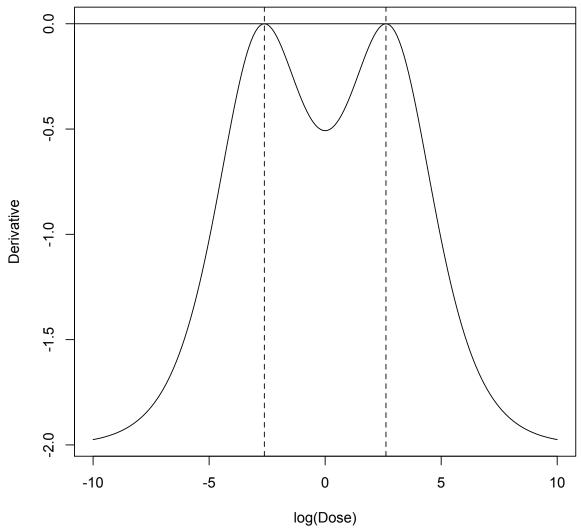

Figure 5.7: Derivative function for $D_{\boldsymbol{\beta}}$-optimal design obtained using the MQL approximation. Vertical dotted lines show the location of the support doses. Note doses are plotted on the log scale.

5.6 Discussion

The numerical construction of D-optimal, including Bayesian D-optimal, multiple dosing designs for the unit variation model is computationally feasible within the restricted class we have outlined. The restricted class in general might not contain the overall D-optimal design, however the restricted designs may be used as a benchmark against which to measure other design strategies.

An issue may arise with the use of asymptotic normal approximations to the distributions of the parameter estimators when the true value of σ^2 is small. Since $\hat{\sigma}^2$ is bounded below by 0, the distribution will be noticably skewed unless the variance is small (which happens only if the sample size is large). In the linear mixed effects model case there is positive probability that $\hat{\sigma}^2 = 0$ which can lead to the asymptotic variance approximation being inaccurate (see McCulloch & Searle, 2001, pp. 39–42). This phenomenon may also be present for the models under consideration here: cases where $\hat{\sigma}^2 = 0$ are mentioned by Xue and Brookmeyer (1997). To produce more accurate measures of estimator variability, it is conceivable that one might wish to look instead at the distribution of $\log \hat{\sigma}^2$ which could plausibly be closer to Gaussian with smaller sample sizes. We do not investigate this in detail here, but we do make some observations. By the chain rule, $\partial P / \partial \log \sigma^2 = \sigma^2 \partial P / \partial \sigma^2$, and so only a minor modification of the computational scheme in Section 5.2.1 is necessary to evaluate $\partial P / \partial \boldsymbol{\theta}$, and therefore also the information matrix, under the new parameterisation.

Another question is how to improve the numerical procedures for ML estimation: at the moment, optimisation of the likelihood takes around 1 minute on a MacBook Pro laptop with a 2.4GHz Intel Core i5 processor. Clearly this is not prohibitive for point estimation, but simulation-based assessments of estimator variability are still fairly involved. The implementation of fast estimation procedures for this model would be helpful in seeing this kind of study design and analysis adopted in practice.

For designs in the restricted class, it is extremely likely that tabulation of elementary integrals similar to those in Section 5.2.1 could be helpful also for estimation. A good starting point for a procedure might be the following, which we have not implemented and do not pursue further. First, construct interpolation tables for a large number of σ^2, say on a finely spaced grid in $[e^{-10}, e^3]$. One should then be able to obtain relatively quickly the estimates of $\boldsymbol{\beta}$ for σ^2 held fixed, which we call $\hat{\boldsymbol{\beta}}(\sigma^2)$. These 'conditional' estimates could be used to compute a profile likelihood for σ^2 evaluated on a grid. A smooth interpolation of these profile likelihood values could then be used to find an approximation to $\hat{\sigma}^2$, and hence $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{\sigma}^2)$.

5.7 Appendix: Differentiation with respect to σ^2

In this section, we show that the application of differentiation under the integral to evaluate $\partial P_t / \partial \sigma^2$ is valid. We consider the uncensored case in detail, the censored case follows an analogous argument. Let $\epsilon > 0$ be arbitrary, $U = (\epsilon, \infty)$, $S = \mathbb{R}$ and $f : U \times S \to \mathbb{R}$ be defined by

$$f(\sigma^2, u) = h(\eta_t + u) \prod_{j=1}^{t-1} \{1 - h(\eta_j + u)\} \phi_{\sigma^2}(u) \,,$$

so that $P_t(\boldsymbol{\eta}, \sigma^2) = \int_{\mathbb{R}} f(\sigma^2, u) d\mu(u)$, where μ denotes Lebesgue measure. Considered as a function of $u \in S$, f is integrable for each $\sigma^2 \in U$, since it is continuous and dominated by $\phi_{\sigma^2}(u)$. Moreover, for fixed $u \in S$, f is differentiable for all $\sigma^2 > 0$, with

$$\begin{aligned}\frac{\partial f}{\partial \sigma^2} &= h(\eta_t + u) \prod_{j=1}^{t-1} \{1 - h(\eta_j + u)\} \frac{\partial \phi_{\sigma^2}}{\partial \sigma^2}(u) \\ &= h(\eta_t + u) \prod_{j=1}^{t-1} \{1 - h(\eta_j + u)\} \frac{1}{2\sigma^2}(1 - \sigma^2)\phi_{\sigma^2}(u) \,.\end{aligned}$$

On $U = (\epsilon, \infty)$, the function $\sigma^2 \mapsto \frac{1}{2\sigma^2}(1 - \sigma^2)$ is bounded above, say by a constant $K > 0$. Therefore also

$$\left| \frac{\partial f}{\partial \sigma^2} \right| \leq K \phi_{\sigma^2}(u) \,,$$

for all $\sigma^2 \in U$ and all $u \in S$, in other words the partial derivative is dominated by an integrable function. Therefore we may apply Theorem 3.1 to obtain that

$$\begin{aligned}\frac{\partial P_t}{\partial \sigma^2}(\boldsymbol{\eta}, \sigma^2) &= \int_{\mathbb{R}} \frac{\partial f}{\partial \sigma^2}(\sigma^2, u) \, d\mu(u) \\ &= \int_{\mathbb{R}} h(\eta_t + u) \prod_{j=1}^{t-1} \{1 - h(\eta_j + u)\} \frac{\partial \phi_{\sigma^2}}{\partial \sigma^2}(u) \, d\mu(u) \,,\end{aligned}$$

for all $\sigma^2 > \epsilon$. However, the choice of $\epsilon > 0$ was arbitrary and so the above holds for all $\sigma^2 > 0$.

Part III

Miscellaneous topics

Chapter 6

Designs for Hierarchical Generalised Linear Models

6.1 Introduction

In Chapters 2 and 3, we developed methodology for deriving efficient designs for GLMMs. The purpose of using such models is to take into account non-normality of the response distribution and correlation between observations within the same block. An alternative modelling strategy would be to use Hierarchical Generalised Linear Models (HGLMs). In this chapter, we develop design methodology for HGLMs.

The family of HGLMs, introduced by Lee and Nelder (1996, 2001), extends GLMMs, primarily by allowing random effects distributions other than the Gaussian. In particular, the use of random effects distributions which are conjugate to the exponential family used for the response can simplify some of the computations involved. The main innovation of the aforementioned papers was to suggest maximisation of the 'h-likelihood' as a technique for the joint estimation of the fixed and random effects. This proposition generated substantial controversy not least in the discussion adjoining Lee and Nelder (1996). The asymptotic properties of the method were studied, and shown to be comparable to marginal likelihood inference under certain regularity conditions. The computational simplifications resulting from the use of h-likelihood make the technique much cheaper to implement than marginal likelihood inference in a GLMM.

We will show in this chapter that the computational advantages of the h-likelihood approach carry over also to the design problem, where they are possibly more pronounced. As a result, the use of approximations to the Fisher information matrix considered in the GLMM context in Chapters 2 and 3 are unnecessary for HGLMs. One of our contributions is the suggestion of an appropriate design optimality criterion, which is based on optimisation of a h-likelihood analogue of *Fisher* information, as opposed to the analogue of the *observed* information which is employed by Lee, Nelder and co-authors for their inferences. For more details, see Section 6.3.

The chapter is organised as follows. In Section 6.2 we define the class of HGLMs and give details of the h-likelihood estimation procedure. Section 6.3 considers different forms of the information matrix for HGLMs, and uses these to motivate various optimality criteria. We consider several different design structures, including split-plot designs, with differing degrees of restrictions of the factors. These structures are outlined, together with corresponding optimisation algorithms in Section 6.4. Examples of the construction of designs using the methods developed are given in Sections 6.5 and 6.6.

6.2 Hierarchical generalised linear models

We focus on HGLMs in which the influence of the blocks appears through a random intercept term. These models are defined as follows. Let y_{ij} denote the response of the jth unit in the ith block, $i = 1, \dots, n_b$, $j = 1, \dots, m_i$, and further let $\mathbf{x}_{ij}$ denote the vector of values, applied to this unit, of the q controllable variables. Let $N = \sum_{i=1}^{n_b} m_i$ be the total number of observations. Associated with the ith block in the experiment there is a corresponding random effect, or random intercept, denoted by v_i.

Conditional on v_i, the responses in block i are independent and follow a generalised linear model. In other words, the conditional distribution of y_{ij} given v_i is an exponential family $\pi(\mu_{ij})$ with mean μ_{ij}, and variance $\phi V(\mu_{ij})$. The mean relates to the controllable variables via the linear predictor η_{ij} and the link function g_μ, as follows:

$$g_\mu(\mu_{ij}) = \eta_{ij} = \mathbf{f}^T(\mathbf{x}_{ij})\,\boldsymbol{\beta} + v_i\,, \tag{6.1}$$

where $\boldsymbol{\beta}$ is the vector of p fixed effects parameters, and $\mathbf{f} : \mathbb{R}^q \to \mathbb{R}^p$ is a known function which specifies which terms are to be included in the linear predictor. For example, $\mathbf{f}(\mathbf{x})$ may contain just first order terms in the entries of $\mathbf{x}$, or it may also include other polynomial terms such as quadratics or interactions.

To complete the definition of the HGLM, we must specify the distribution of the response, and the random effects. Unlike in a GLMM, the specification of the distribution of v_i is not done directly, instead we first relate v_i to a random effect u_i on a different scale, via a link function g_r for the random effects. Thus $g_r(u_i) = v_i$. Then the distribution of u_i is specified, with a (vector) parameter $\boldsymbol{\alpha}$, to determine the model. For examples of choices of distributions and link functions to determine the HGLM, see Table 6.1. The random effects distributions used in HGLMs commonly have two parameters. To ensure identifiability, Lee and Nelder (1996) recommended imposing a constraining relation on the parameters to yield a one-parameter distribution. Suggested restrictions are also listed in Table 6.1.

Model	Distribution of u	Mode of v	Restriction	Mode after restriction	Link ($g_r = g_\mu$)
Poisson-gamma	Gamma(α_1, α_2)	$\log(\alpha_1\alpha_2)$	$\alpha_1\alpha_2 = 1$	0	log
Binomial-beta	Beta(α_1, α_2)	$\log\left(\frac{\alpha_1-1}{\alpha_2-1}\right)$	$\alpha_1 = \alpha_2$	0	logit
Normal-normal	Normal(α_1, α_2)	α_1	$\alpha_1 = 0$	0	identity

Table 6.1: Choices of distribution and link function in HGLMs. In each case the first part of the model name defines the conditional distribution of the response, and the usual corresponding GLM variance function is used. The second part of the name gives the distribution of u_i.

An equivalent vector statement of (6.1) can be obtain by writing the data in 'long' format and defining appropriate model matrices. These model matrices will be useful later when stating the form of the information matrix. Let us denote by $\mathbf{y}$ the vector of responses y_{ij} written in lexicographical order, grouped by block, i.e.

$$\mathbf{y} = (y_{11}, \ldots, y_{1m_1}, y_{21}, \ldots, y_{2m_2}, \ldots, y_{n_b 1}, \ldots, y_{n_b m_{n_b}})^T .$$

Let us also denote by F be the fixed effects model matrix whose rows are the $\mathbf{f}^T(\mathbf{x}_{ij})$, again in lexicographical order. Finally let us write Z for the $N \times n_b$ indicator matrix that identifies the block to which an observation belongs, that is

$$Z = \begin{pmatrix} \mathbf{1}_{m_1} & & & 0 \\ & \mathbf{1}_{m_2} & & \\ & & \ddots & \\ 0 & & & \mathbf{1}_{m_{n_b}} \end{pmatrix},$$

with $\mathbf{1}_k = (1, 1, \ldots, 1)^T$ a column vector consisting of k ones. Then the model equation (6.1) can be restated as

$$g_\mu(\mathrm{E}(\mathbf{y}|\mathbf{v})) = F\boldsymbol{\beta} + Z\mathbf{v} ,$$

where g_μ acts elementwise and $\mathbf{v} = (v_1, \ldots, v_{n_b})^T$.

6.2.1 *h*-likelihood inference

In a series of papers (Lee and Nelder, 1996, 2001, 2009; Lee, Nelder and Noh, 2007), Lee, Nelder and co-authors advocate the use of h-likelihood as a device for the joint estimation of fixed and random effects. The h-(log)likelihood is defined as

$$h(\boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{v}; \mathbf{y}) = \log f_{\mathbf{y}|\mathbf{v}}(\mathbf{y}|\mathbf{v}; \boldsymbol{\beta}, \boldsymbol{\alpha}) + \log f_{\mathbf{v}}(\mathbf{v}; \boldsymbol{\alpha}) , \tag{6.2}$$

considered as a function of both the fixed and random effects, $\mathbf{v}$ and $\boldsymbol{\beta}$. In (6.2), $f_{\mathbf{y}|\mathbf{v}}$ and $f_{\mathbf{v}}$ denote respectively (i) the conditional density of the responses given the random effects, and (ii) the density function of the random effects $\mathbf{v}$. Thus the h-likelihood is formed by taking the joint density function of the data and the random effects, and viewing it as a function of $\mathbf{v}$ and $\boldsymbol{\beta}$. Clearly (6.2) is not an orthodox likelihood as $\mathbf{v}$ is an unobservable random variable. The maximum h-likelihood estimators are given by

$$(\hat{\boldsymbol{\beta}}^T, \hat{\mathbf{v}}^T)^T = \underset{\boldsymbol{\beta}, \mathbf{v}}{\arg\max}\; h(\boldsymbol{\beta}, \boldsymbol{\alpha}\mathbf{v}; \mathbf{y}) .$$

Note that maximum h-likelihood estimation of $(\boldsymbol{\beta}^T, \mathbf{v}^T)^T$ is equivalent to Bayesian maximum a posteriori estimation of $(\boldsymbol{\beta}^T, \mathbf{v}^T)$ with an improper uniform prior on $\boldsymbol{\beta}$ (see for example the comment by D. Clayton in the discussion following Lee and Nelder, 1996). In this chapter we assume that the parameters, $\boldsymbol{\alpha}$, of the random effects distribution are known.

Much of the controversy surrounding the use of h-likelihood appeared to stem from the suggestion of Lee and Nelder (1996) to use h-likelihood estimates also in the case when interest lies only in $\boldsymbol{\beta}$, rather than the usual marginal likelihood,

$$\ell(\boldsymbol{\beta}; \mathbf{y}) = \int f(\mathbf{y}, \mathbf{v}; \boldsymbol{\beta}, \boldsymbol{\alpha})\, d\mathbf{v} .$$

This idea was later revised: Lee and Nelder (2009, Section 3.2) suggest using the marginal likelihood, but interpret the 'marginal likelihood as an adjusted profile likelihood (in the sense of Barndorff-Nielsen, 1983) eliminating nuisance unobservables $\mathbf{v}$ from the h-likelihood'. Thus they regard the h-likelihood as being more fundamental than the marginal likelihood but recognise the need to eliminate nuisance parameters. The HGLM literature also proposes techniques for estimation of $\boldsymbol{\alpha}$, by a restricted likelihood conditional on the marginal estimates of $\boldsymbol{\beta}$, but we do not consider this aspect here. The limitation of scope not to consider the quality of estimation of variance components is also present in work on design for linear mixed models, e.g. Goos and Vandebroek (2001), and is an avenue for future work.

In this chapter we take an agnostic stance with regard to abstract inferential principles, and evaluate h-likelihood estimators from a straightforward frequentist viewpoint. In particular, in Appendix 6.9 we find that the asymptotic approximations to the variance of $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{v}})$ given in Lee and Nelder (1996) seem to be accurate enough to be used as the basis for optimal designs. We found the asymptotic approximations to be accurate even when there are restrictions on the design structure such that fixed effects estimation of the block effects would be impossible, for instance when we must use a split-plot design.

6.3 Optimality criteria

6.3.1 Information matrices

Before we define our optimality criteria, we first discuss several asymptotically equivalent expressions which allow us to approximate the variance of h-likelihood estimators of the model parameters. This offers us several potential generalisations of the classical information matrix. We will base our design optimality criteria on one of these generalisations, the 'marginal expected h-information matrix', a choice which is justified below. In this section, we use ξ to denote the (exact) design of the experiment, which is defined by the $\mathbf{x}_{ij}$, $1 \leq i \leq n_b$, $1 \leq j \leq m_i$.

Let us define H to be the negative Hessian matrix of the h-(log)likelihood,

$$H(\xi; \boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{v}, \mathbf{y}) = \begin{pmatrix} -\frac{\partial^2 h}{\partial \boldsymbol{\beta}^2} & -\frac{\partial^2 h}{\partial \boldsymbol{\beta} \partial \mathbf{v}^T} \\ -\frac{\partial^2 h}{\partial \mathbf{v} \partial \boldsymbol{\beta}^T} & -\frac{\partial^2 h}{\partial \mathbf{v}^2} \end{pmatrix} .$$

We define the *observed h-information matrix* as the estimate of H,

$$\hat{H} = H(\xi; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\mathbf{v}}, \mathbf{y}_{\text{obs}}) ,$$

where $\mathbf{y}_{\text{obs}}$ is an observed vector of responses. In the case where $\boldsymbol{\alpha}$ is known, clearly we may use $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}$. The *conditional expected h-information* is defined to be

$$J_C(\xi; \boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{v}) = \mathrm{E}_{\mathbf{y}|\mathbf{v}}[H(\xi; \boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{v}, \mathbf{y})|\mathbf{v}] ,$$

which can be estimated from data as $\hat{J}_C = J_C(\xi; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\mathbf{v}})$. Finally we also define the *marginal expected h-information* as

$$\begin{aligned} J_M(\xi; \boldsymbol{\beta}, \boldsymbol{\alpha}) &= \mathrm{E}_{\mathbf{y},\mathbf{v}}[H(\xi; \boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{v}, \mathbf{y})] \\ &= E_{\mathbf{v}}[J_C(\xi; \boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{v}))] . \end{aligned}$$

Defined thus, J_M can be estimated from data as $\hat{J}_M = J_M(\xi; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$. In the above definitions $\mathrm{E}_{\mathbf{y}|\mathbf{v}}$ denotes an average with respect to the conditional distribution of $\mathbf{y}$ given $\mathbf{v}$, whereas $\mathrm{E}_{\mathbf{y},\mathbf{v}}$ denotes expectation with respect to the joint distribution of $\mathbf{y}$ and $\mathbf{v}$. Note that we refer to J_M as the 'marginal information matrix' because it is the marginal expectation of the negative Hessian, as opposed to J_C which is a conditional mean.

Lee and Nelder (1996) refer, somewhat implicitly, to some of the above matrices as asymptotic approximations to the (inverse) marginal variance of the error in the h-likelihood estimators, particularly in the case where $\boldsymbol{\alpha}$ is known. In other words,

$$\mathrm{var}\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{v}} - \mathbf{v} \end{pmatrix} \approx \hat{H}^{-1},\, \hat{J}_C^{-1},\, J_C^{-1},\, \hat{J}_M^{-1},\, J_M^{-1}\,. \tag{6.3}$$

They make use of $\hat{J}_C^{-1}$ when performing inference on $\boldsymbol{\beta}$, $\mathbf{v}$. For the observed, conditional expected and marginal expected information see Lee and Nelder (1996), Sections 3.3, 4.1 and Appendix C respectively.

Note that the variance in (6.3) is not a conditional variance (for instance upon $\mathbf{v}$). It is a marginal variance, which includes variation arising from the fact that if we were to repeat the experiment several times the block effects would be different each time. As a result, it seems most appropriate at the design phase to use J_M^{-1} as an approximation to the variance, since this does not require us to assume (or estimate) a value of the random effects $\mathbf{v}$ prior to running the experiment. In Section 6.9 we empirically evaluate the use of J_M^{-1} as a variance approximation.

The conditional h-information can be evaluated as (see Lee and Nelder, 1996, Section 4.1, but note the slightly different definition of U)

$$J_C(\xi; \boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{v}) = \frac{1}{\phi}\begin{pmatrix} F^T W F & F^T W Z \\ Z^T W F & Z^T W Z + \phi U \end{pmatrix},$$

where $W = W(\boldsymbol{\beta}, \xi, \mathbf{v})$ is the diagonal matrix of GLM weights with diagonal entries

$$\left(\frac{\partial \mu_{ij}}{\partial \eta_{ij}}\right)^2 V(\mu_{ij})^{-1} = \frac{1}{[g'_\mu(\mu_{ij})]^2 V(\mu_{ij})}\,, \tag{6.4}$$

written in lexicographical order. The matrix U is diagonal with ith entry $-\partial^2 \log f_{\mathbf{v}}(\mathbf{v}; \boldsymbol{\alpha})/\partial v_i^2$. The matrices F and Z are as given in Section 6.2. The marginal h-information, on which we base optimal designs, can be evaluated as

$$J_M(\xi; \boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{1}{\phi}\begin{pmatrix} F^T\,\mathrm{E}(W)F & F^T\,\mathrm{E}(W)Z \\ Z^T\,\mathrm{E}(W)F & Z^T\,\mathrm{E}(W)Z + \phi\,\mathrm{E}(U) \end{pmatrix}, \tag{6.5}$$

where the expectations here are with respect to $\mathbf{v}$. These expectations can be evaluated using the details given in the appendix, Section 6.8.

6.3.2 Optimality criteria

Clearly, in common with the situation for other complex models, optimal designs for HGLMs will depend on the values of the parameters $(\boldsymbol{\beta}, \boldsymbol{\alpha})$. In view of this, we say an exact design ξ^* is *locally D-optimal at* $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ if it maximises

$$\psi_D(\xi; \boldsymbol{\beta}, \boldsymbol{\alpha}) = |J_M(\xi; \boldsymbol{\beta}, \boldsymbol{\alpha})|,$$

for particular assumed values of the parameters.

Sometimes interest may lie in the fixed effects parameters only. We will therefore also consider D_S-optimal designs (Atkinson et al., 2007, p. 138) for estimating $\boldsymbol{\beta}$ as precisely as possible, again based on J_M. Specifically, a design ξ^* is *locally D_S-optimal for $\boldsymbol{\beta}$* (at $\boldsymbol{\beta}, \boldsymbol{\alpha}$) if it minimises

$$\psi_{D_S}(\xi; \boldsymbol{\beta}, \boldsymbol{\alpha}) = |\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\xi; \boldsymbol{\beta}, \boldsymbol{\alpha})|,$$

where $V_{\boldsymbol{\beta}\boldsymbol{\beta}}(\xi; \boldsymbol{\beta}, \boldsymbol{\alpha})$ is the top-left $p \times p$ (i.e. $\boldsymbol{\beta}$) part of $J_M^{-1}(\xi; \boldsymbol{\beta}, \boldsymbol{\alpha})$. This can be evaluated in the following way: decomposing J_M into block submatrices as

$$J_M = \begin{pmatrix} M_{\boldsymbol{\beta}\boldsymbol{\beta}} & M_{\boldsymbol{\beta}\mathbf{v}} \\ M_{\mathbf{v}\boldsymbol{\beta}} & M_{\mathbf{v}\mathbf{v}} \end{pmatrix},$$

where $M_{\boldsymbol{\beta}\boldsymbol{\beta}}$ is $p \times p$ and $M_{\mathbf{v}\mathbf{v}}$ is $n_b \times n_b$, we have that (Atkinson et al., 2007)

$$V_{\boldsymbol{\beta}\boldsymbol{\beta}} = \{M_{\boldsymbol{\beta}\boldsymbol{\beta}} - M_{\boldsymbol{\beta}\mathbf{v}} M_{\mathbf{v}\mathbf{v}}^{-1} M_{\mathbf{v}\boldsymbol{\beta}}\}^{-1},$$

and so an equivalent criterion for ξ^* to be D_S optimal is that

$$\Phi_{D_S}(\xi; \boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{1}{\phi}|F^T \operatorname{E}(W)F - F^T \operatorname{E}(W)Z[Z^T \operatorname{E}(W)Z + \phi \operatorname{E}(U)]^{-1} Z^T \operatorname{E}(W)F|$$

is maximised at $\xi = \xi^*$. The D_S criterion takes into account that we are also estimating the random effects but does not make precise estimation of $\mathbf{v}$ a consideration, other than insofar as it affects estimation of $\boldsymbol{\beta}$.

Note that if $\mathbf{v}$ is truly nuisance then, as was stated in Section 6.2.1, strictly we should use marginal likelihood estimation of, and inference about, $\boldsymbol{\beta}$. Thus the use of such D_S-optimal designs is not totally principled. However, D_S-optimal designs do provide an interesting point of comparison for the D-optimal GLMM designs of Chapters 2 and 3. Recall that in these chapters, we computed D-optimal designs for GLMMs using various approximations to the Fisher information matrix associated with maximum (marginal) likelihood estimation.

In order to obtain designs which are more robust to prior uncertainty about the values of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$, we will also compute pseudo-Bayesian D-optimal designs, in other words designs ξ which maximise the objective function of Chaloner and Larntz (1989), i.e.

$$\psi_{\text{Bayes}}(\xi; f) = \int \log \psi(\xi; \boldsymbol{\beta}, \boldsymbol{\alpha})\, f(\boldsymbol{\beta}, \boldsymbol{\alpha})\, d\boldsymbol{\beta}\, d\boldsymbol{\alpha}, \tag{6.6}$$

where f is a prior density function on $(\boldsymbol{\beta}, \boldsymbol{\alpha})$. There is not an assumption that the resulting data analysis will be Bayesian, this 'prior' is merely a device to ensure a greater degree of robustness to a range of possibilities for the true values of the parameters. As discussed in Chapter 7, the objective function (6.6) may fail to converge if the support of the prior distribution is too large or contains singularities. For the sake of simplicity, here we restrict our attention to uniform priors that avoid singularities.

6.4 Design structures

In much experimental design literature the objects of focus are *continuous designs*, in other words designs defined through a finitely supported probability measure which is independent of the sample size. Such a measure represents the fraction of available resource that should be allocated to particular experimental conditions. Continuous designs do not make sense in the context of h-likelihood estimation, since the the following quantities depend on the sample size: (i) the number of random effects parameters, (ii) the number of arguments in the h-likelihood, and (iii) the dimension of the information matrix. Therefore we focus exclusively on exact designs where the sample size is fixed. Within the exact design framework, we consider various design structures corresponding to different degrees of restriction on the factors in the experiment.

A *split-plot* experiment is a blocked experiment in which one or more of the factors are restricted to have the same value for all runs in each block. For a discussion of the merits of such designs in the context of industrial experiments see Jones and Nachtsheim (2009). To properly analyse the data from these experiments, mixed models are necessary. The restricted factors in a split-plot experiment are referred to as *whole-plot* factors, and the remaining free factors are referred to as *sub-plot* factors. Whole-plot factors may also be referred to as 'hard-to-change', with a typical example being the temperature of an oven.

In this chapter we derive designs of three types: unrestricted, split-plot, and 'whole-plot'. By a whole-plot design we mean a split-plot design in which all of the factors are whole-plot factors (and so there are no sub-plot factors). For an instance of a whole-plot design, see that used in the count-response wave-solder experiment reported by Hamada and Nelder (1997), or the binomial-response seed germination experiment discussed by Breslow and Clayton (1993). The first pair of authors analysed their data using overdispersed fixed-effects generalised linear models, which model the extra variation introduced by the presence of blocks by including a single extra parameter. However, these models are unable to take into account the correlation between observations in the same block which is introduced by the block effects if they are present. Also, in the fixed-effects framework it is impossible to fit a separate parameter for each block, as a result of the total confounding of all of the factors with blocks. A mixed effects model analysis of these types of experiments, for example using HGLMs, would allow the consideration of block effects. This is indeed the approach taken in the second example by Breslow and Clayton (1993), who modelled their data using GLMMs. We consider the wave-solder experiment further in Section 6.5.

6.4.1 Algorithms

To find optimal unrestricted, split-plot, and whole-plot designs we use a co-ordinate optimisation algorithm similar to the 'candidate-set-free' approach of Jones and Goos (2007). Let f_w and f_s denote the number of whole-plot and sub-plot factors respectively. Throughout the algorithm, the computer holds two arrays in memory: an $n_b \times f_w$ whole-plot factor array, and an $N \times f_s$ sub-plot factor array. In the sub-plot factor array, the first m_1 rows correspond to the factor values used in the first block, the next m_2 rows to the second block, and so on. In the whole plot factor array, the ith row corresponds to the values of the whole plot factors used in the ith block. For an illustration of the setup of these arrays, see Tables 6.2 and 6.3, in which $W_1, \dots, W_{f_w}$

represent the whole-plot factors and $S_1, \dots, S_{f_s}$ represent the sub-plot factors.

The algorithm begins by generating random designs until a non-singular design is found. The main loop of the algorithm consists of repeatedly performing *passes*. In each pass of the algorithm, we optimise each element of the array in 'typewriter fashion' moving across each row from left to right. Upon reaching the end of a row we move to the leftmost entry in the next row. This is done first for the whole-plot factor array, and second for the sub-plot factor array.

For each element, all possible changes its value are considered (the factors are discrete in this work). The value of the objective function is calculated for each proposed update to the co-ordinate. The change which would maximise the objective function value is kept, and we then move on to the next element in the array. The algorithm terminates after a complete pass yields no changes. Since the algorithm is greedy, it is prone to becoming stuck in sub-optimal attractor states. To mitigate this, the best design from multiple random initialisations is chosen.

When finding unrestricted designs, there is no whole-plot array, as there are no whole-plot factors. Conversely, when finding whole-plot designs there is no sub-plot array.

Block	W_1	W_2	$\dots$	W_{f_w}
1	$\bullet$	$\bullet$	$\dots$	$\bullet$
2	$\bullet$	$\bullet$	$\dots$	$\bullet$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n_b	$\bullet$	$\bullet$	$\dots$	$\bullet$

Table 6.2: Whole-plot factor array used in the co-ordinate optimisation algorithm. Large dots, $\bullet$, represent arbitrary values of the factors.

	Unit	S_1	S_2	$\dots$	S_{f_s}
Block 1	$(1,1)$	$\bullet$	$\bullet$	$\dots$	$\bullet$
	$(1,2)$	$\bullet$	$\bullet$	$\dots$	$\bullet$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	$(1,m_1)$	$\bullet$	$\bullet$	$\dots$	$\bullet$
Block 2	$(2,1)$	$\bullet$	$\bullet$	$\dots$	$\bullet$
	$(2,2)$	$\bullet$	$\bullet$	$\dots$	$\bullet$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	$(2,m_2)$	$\bullet$	$\bullet$	$\dots$	$\bullet$
$\vdots$	$\vdots$			$\vdots$	
Block n_b	$(n_b,1)$	$\bullet$	$\bullet$	$\dots$	$\bullet$
	$(n_b,2)$	$\bullet$	$\bullet$	$\dots$	$\bullet$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	(n_b,m_{n_b})	$\bullet$	$\bullet$	$\dots$	$\bullet$

Table 6.3: Sub-plot factor array used in the co-ordinate optimisation algorithm. Large dots, $\bullet$, represent arbitrary values of the factors.

6.5 Example: wave-solder experiment

Hamada and Nelder (1997) discuss an experiment, reported by Condra (1993), investigating a

	Factor							y		
Block	A	B	C	D	E	F	G	1	2	3
1	0	0	0	0	0	0	0	13	30	26
2	0	0	0	1	1	1	1	4	16	11
3	0	0	1	0	0	1	1	20	15	20
4	0	0	1	1	1	0	0	42	43	64
5	0	1	0	0	1	0	1	14	15	17
6	0	1	0	1	0	1	0	10	17	16
7	0	1	1	0	1	1	0	36	29	53
8	0	1	1	1	0	0	1	5	9	16
9	1	0	0	0	1	1	0	29	0	14
10	1	0	0	1	0	0	1	10	26	9
11	1	0	1	0	1	0	1	28	173	19
12	1	0	1	1	0	1	0	100	129	151
13	1	1	0	0	0	1	1	11	15	11
14	1	1	0	1	1	0	0	17	2	17
15	1	1	1	0	0	0	0	53	70	89
16	1	1	1	1	1	1	1	23	22	7

Table 6.4: Fractional factorial design used in the wave soldering experiment from Hamada and Nelder (1997), with response data

wave-soldering process for electronic circuit card assembly. We use this example to motivate the construction of more efficient designs for similar experiments with larger blocks.

In the original experiment there were 7 factors, labelled A–G, each with two levels coded by 0/1 as is common in generalised linear model analyses. We opt to keep this coding in what follows, contrary to the coding of ±1 usual in the design literature, since we prefer consistency with the original example. The response was the number of defects per plate, and the data were analysed with an overdispersed Poisson model. The effects of interest were the main effects of A–G, together with six two-factor interactions between A–D, namely AB, AC, AD, BC, BD and CD. The aim of the study was to discover which combination of factor levels minimised the average number of defects.

The design actually used had its treatments taken from a 2^{7-3} fractional factorial design (see Table 6.4), and the plates were soldered in 16 batches of 3 plates. Within any given batch, the same process settings were used for all three plates. We argue that this potentially constitutes a whole-plot block structure.

We re-analysed the original data set using a Poisson-gamma HGLM. The terms we chose to include in the linear predictor were the same as in the final GLM model in Hamada and Nelder (1997), i.e. the main effect of F was not included and the only interactions fitted were AC and BD. The `R` package `HGLMMM` (Molas and Lesaffre, 2011) was used to estimate the fixed effects parameters and variance component only, as the particular values of the random effects are not relevant in future experiments. The option to use an approximation to the marginal likelihood rather than the h-likelihood was selected because this is recommended when interest is in the fixed effects only (Lee et al., 2007). The parameter estimates obtained are given are given in Table 6.5.

We computed a locally D-optimal whole-plot design for estimating the HGLM, using these refitted parameter values. The factor settings for this design are given in Table 6.6. We assumed that the batches in the future experiment would be of size 10, rather than 3, in order for the asymptotic variance approximations to hold reasonably well.

In the fractional factorial design shown in Table 6.4, each factor has the property that the high and low levels are used the same number of times. We refer to this as property as *balance (in the factors)*. In contrast, the locally D-optimal design is substantially unbalanced. In particular, factor G is set to its high level only for only 1/16 of the available runs, in other words around 6% of the time. This lack of balance property is a recurring theme in D-optimal designs for Poisson-gamma HGLMs, for additional examples see Section 6.6.2. The efficiency gain in using the locally optimal design is fairly substantial: the efficiency of the design actually used was 81.3%.

Term	Estimate	Std. Error
Intercept	3.101	0.157
A	-0.144	0.159
B	0.301	0.149
C	0.472	0.151
D	0.523	0.146
E	-0.194	0.106
G	-0.757	0.107
AC	0.726	0.213
BD	-1.215	0.213
$\alpha_2 = \text{var}(u_i)$	0.027	-

Table 6.5: Parameter estimates for the refitted HGLM. The parameter α_2 denotes the scale parameter of the gamma distribution for u_i.

	Factor						
Block	A	B	C	D	E	F	G
1	0	0	0	1	0	1	0
2	0	0	1	0	0	0	0
3	0	0	1	1	1	0	0
4	0	1	0	0	1	0	0
5	0	1	0	1	0	0	0
6	0	1	1	0	0	1	0
7	1	0	0	0	0	0	0
8	1	0	0	1	1	0	0
9	1	0	1	0	1	1	0
10	1	0	1	1	0	0	0
11	1	0	1	1	0	0	0
12	1	0	1	1	0	1	0
13	1	0	1	1	0	1	1
14	1	1	0	0	0	1	0
15	1	1	1	0	0	0	0
16	1	1	1	1	1	1	0

Table 6.6: Locally D-optimal wholeplot HGLM design for the wave soldering experiment, assuming runs of size 10

6.5.1 Lack of balance

In the wave-solder example, we observed that the D-optimal design can be quite unbalanced. That such a degree of imbalance is optimal is perhaps at first surprising. However, similar phenomena have been observed before in the context of Poisson designs without random effects. In particular, Russell, Woods, Lewis and Eccleston (2009) derive an analytical form for the locally D-optimal approximate design in the first-order Poisson regression model, under certain restrictions on the parameter values. The covariates are assumed to be continuous. In the resulting optimal designs, each of the variables is set to only two levels, one of which is at the end of the allowable range and one of which is in the interior. The interior value is used for only $100/(p+1)\%$ of the runs in the experiment, where p is the number of factors (equivalently, the number of non-intercept fixed effects parameters). Thus, D-optimal Poisson designs can be quite unbalanced. Note that if β_j is positive, then the majority of the time x_j is set to the highest possible level, whereas if $\beta_j < 0$ then x_j is most often set to the lowest possible level.

The key to understanding this feature is to consider the heteroscedasticity assumption in the Poisson model. Settings of the factors, $\mathbf{x}$, which have a higher mean response also have a higher variance. Thus to be able to estimate with high precision the mean at such an $\mathbf{x}$, one needs to assign more experimental effort here than to settings where the mean (and so too the variance) is lower.

A numerical study was conducted to investigate the sensitivity of balance in the D-optimal HGLM design to the values of the parameters. The value of the intercept and main effects parameters for A–G were varied one at a time, holding all other parameters constant. In each case, the range tried was $[-1.5, 1.5]|\tilde{\beta}_i|$, where $\tilde{\beta}_i$ is the corresponding value estimated from the original data set (except for F, where a range of [-3,3] was used). The value of β_0 was found not to affect the optimal design.

The results of the numerical study are shown in Figure 6.1. Each of the panels in this figure corresponds to one of the factors A-G. The horizontal axis of the ith panel shows the value of the main effect parameter, β_i, for the corresponding factor. The vertical axis on the ith panel shows the proportion of runs in the experiment which have $x_i = 1$. High or low proportions correspond to a lack of balance, whilst those near 0.5 indicate near perfect balance. In each case, the original fitted parameter value $\tilde{\beta}_i$ is indicated by a vertical dotted line. It is clear from the figure that as β_i increases, a greater proportion of runs use $x_i = 1$. The parameter values for the non-interacting factors E, F, G have the greatest impact on the optimal design. However, the sizes of the effects of factors C and D are also important.

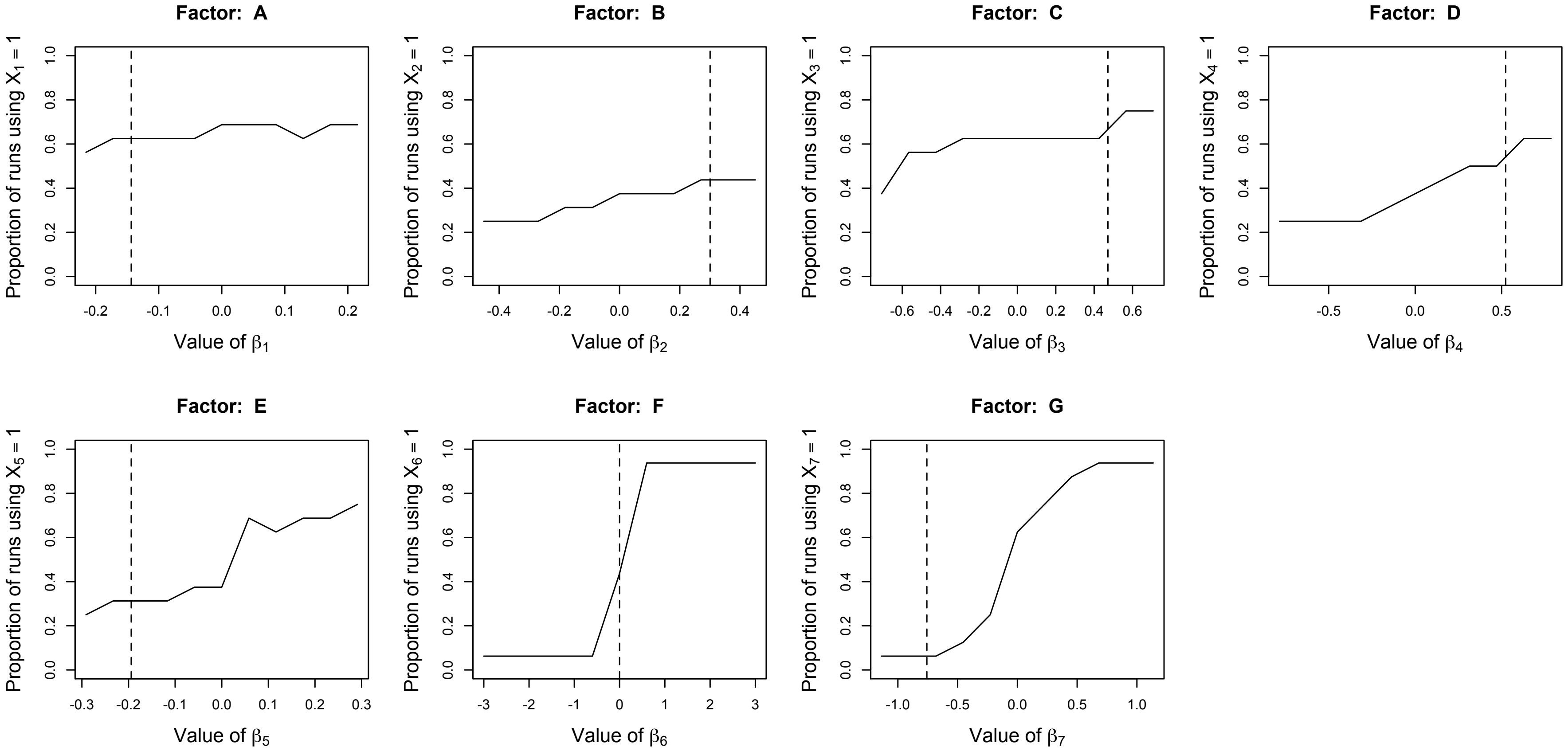

Figure 6.1: Locally D-optimal designs for estimating Poisson-gamma HGLM in wave solder experiment: sensitivity of balance to parameter values. Each panel corresponds to a factor whose main effect parameter is varied.

The computations to produce Figure 6.1 are of the ‘overnight’ order: for each of 8 parameters, 11 values were used. For each of these values, 75 random starts of the co-ordinate exchange algorithm were tried. The entire set of optimisations for each parameter value takes around 5-10 minutes on a Macbook Pro computer with a 2.4GHz Intel Core i5 processor.

6.6 Comparison of methods

6.6.1 Alternative approaches

As stated in the introduction, there are alternative conditional modelling strategies which can be used to analyse the data from experiments with non-normal, blocked responses. We will compare the HGLM designs obtained in this chapter to

1. designs for corresponding GLMMs, arising from an MQL approximation to the Fisher information matrix associated with maximum (marginal) likelihood along the lines of Chapter 2.

2. designs for quasi-likelihood estimation of corresponding GLMMs, along the lines of Niaparast (2009). For further background to this approach, see Section 2.6.

We obtain a GLMM corresponding to a particular HGLM by using the same linear predictor structure, and same fixed effects parameter values. The value of the GLMM block effect variance, σ^2, is chosen to make the resulting random effects distribution close to that under the HGLM. For instance, in a Poisson-gamma HGLM, the distribution of $v_i = \log u_i$ is reasonably close to normal with mean 0. A reasonable choice of σ^2 is therefore the variance of v_i.

For the Poisson response GLMM with log-link, the expressions for the information matrices are

$$\begin{aligned} J_{\mathrm{MQL}}(\xi; \boldsymbol{\beta}, \sigma^2) &= \sum_{i=1}^{n_b} F_i^T W_{i,\mathrm{MQL}}^{-1} F_i \\ J_{\mathrm{QL}}(\xi; \boldsymbol{\beta}, \sigma^2) &= \sum_{i=1}^{n_b} F_i^T W_{i,\mathrm{QL}}^{-1} F_i \,, \end{aligned}$$

where F_i is the $m_i \times p$ model matrix for the ith block, and

$$\begin{aligned} W_{i,\mathrm{MQL}} &= \mathrm{diag}(e^{-\mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta}} \,:\, 1 \le j \le m_i) + \sigma^2 \mathbf{1}_{m_i} \mathbf{1}_{m_i}^T \\ W_{i,\mathrm{QL}} &= \mathrm{diag}(e^{-\mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta}} \,:\, 1 \le j \le m_i) + (e^{\sigma^2} - 1) \mathbf{1}_{m_i} \mathbf{1}_{m_i}^T \,, \end{aligned}$$

where $\mathbf{f}$ and $\boldsymbol{\beta}$ are as defined in (6.1), and diag$(\cdot)$ denotes the diagonal matrix with diagonal entries equal to those given in the parentheses. Finally, $\mathbf{1}_{m_i} = (1, 1, \ldots, 1)^T$ is an m_i-vector consisting of ones.

Note that it is much easier to evaluate the marginal expected h-information than it is to compute the GLMM information matrix via the complete enumeration method detailed in Chapters 2 and 3. This is essentially because for the former we do not need to evaluate a separate integral for each possible outcome in the block. Moreover, for the Poisson-gamma HGLM, all necessary integrals to compute the marginal h-information can be evaluated analytically: for details, see Section 6.8.

6.6.2 Four factor example

As another example we consider designs where there are four controllable factors $x_1, \ldots, x_4$ each with two levels, again coded as 0/1. A first order predictor structure is assumed, in other words

$$\eta_{ij} = \beta_0 + \beta_1 x_1^{(ij)} + \beta_2 x_2^{(ij)} + \beta_3 x_3^{(ij)} + \beta_4 x_4^{(ij)} + v_i\,, \tag{6.7}$$

where $x_k^{(ij)}$ is defined by $\mathbf{x}_{ij} = (x_1^{(ij)}, \ldots, x_4^{(ij)})^T$ for $k = 1, \ldots, 4$, $i = 1, \ldots, n_b$, $j = 1, \ldots, m_i$. In this case, $\mathbf{f}(\mathbf{x}) = (1, \mathbf{x}^T)^T$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)^T$.

We shall obtain D- and D_S-optimal designs for the Poisson-gamma HGLM and compare these to MQL and QL designs for a corresponding Poisson GLMM, as described in Section 6.6.1. Locally optimal designs of these types are computed under the parameter scenarios (a)–(d) outlined in Table 6.7. A complete list of the designs calculated to form the evidence for this Section is given in Table 6.8, in terms of the combinations of criterion, design structure and parameter scenario which define the design.

Scenario	$\boldsymbol{\beta}^T$
(a)	$(0, 2, 2, 2)$
(b)	$(0, 0.1, 0.2, 0.3, 0.5)$
(c)	$(0, 0.2, 0.2, 0.2, 0.2)$
(d)	$(0, 0, 0, 0, 0)$

Table 6.7: Parameter scenarios in the four-factor example.

Criterion	Wholeplot	Split-plot	Unrestricted
D-HGLM	b, a^*,c^*,d^*	a, b, c, d	a, b, c, d
D_S-HGLM			b, c, d
MQL-GLMM	a^*,b^*,c^*,d^*	b, c, d, a^*	a, b, c, d
QL-GLMM		b, c, d, a^*	

Table 6.8: Complete list of designs calculated for Section 6.6.2. This table states which parameter scenarios, as defined in Table 6.7, were used to calculate locally optimal designs, cross-classified by criterion and design structure. Starred, italicised entries correspond to designs which were used only to inform the argument, and which are not presented, or used in calculations, in Section 6.6.2.

Throughout, we assume $(\alpha_1, \alpha_2) = (10, 0.1)$ which satisfies the restriction $\alpha_1\alpha_2 = 1$ made in Table 6.1. By simulation, we found that the variance of $v_i = \log u_i$ is approximately 0.1 and so set $\sigma^2 = 0.1$ to obtain an approximating GLMM. We consider the unrestricted, whole-plot and split-plot design structures described in Section 6.4.

Comparison of D-optimal HGLM designs and GLMM designs

The overarching picture for unrestricted and wholeplot designs seems to be that HGLM D-optimal designs tend to replicate treatments with higher means, or equivalently variances, much more strongly than GLMM designs. The D-optimal HGLM and GLMM unrestricted designs under parameter scenario (b) are given (across two tables each) in Tables 6.9–6.10 and 6.11–6.12. In these representations, the first table gives the treatments to be used in terms of their factor settings, and the second table gives the incidence of these treatments in each block. In particular, the larger numbers in the leftmost incidence column in Table 6.10 show that the 'all-high' treatment is replicated quite heavily. Figure 6.2 shows the relationship between the treatment mean and its replication, for HGLM and GLMM unrestricted designs under parameter scenarios (a)–(c). From this figure, we see that there is a strong positive association between the mean response and replication in the leftmost column, corresponding to the HGLM designs. The same trend is not evident in the GLMM column.

	x_1	x_2	x_3	x_4
$\mathbf{t}_1$	1	1	1	1
$\mathbf{t}_2$	1	1	1	0
$\mathbf{t}_3$	1	1	0	1
$\mathbf{t}_4$	1	1	0	0
$\mathbf{t}_5$	1	0	1	1
$\mathbf{t}_6$	1	0	1	0
$\mathbf{t}_7$	1	0	0	1
$\mathbf{t}_8$	0	1	1	1
$\mathbf{t}_9$	0	1	1	0
$\mathbf{t}_{10}$	0	1	0	1
$\mathbf{t}_{11}$	0	0	1	1
$\mathbf{t}_{12}$	0	0	0	1

Table 6.9: D-optimal HGLM unrestricted design under parameter scenario (b). Table shows factor settings for the treatments, $\mathbf{t}_1$–$\mathbf{t}_{12}$, used in the design defined in Table 6.10.

	Treatment												
Block	$\mathbf{t}_1$	$\mathbf{t}_2$	$\mathbf{t}_3$	$\mathbf{t}_4$	$\mathbf{t}_5$	$\mathbf{t}_6$	$\mathbf{t}_7$	$\mathbf{t}_8$	$\mathbf{t}_9$	$\mathbf{t}_{10}$	$\mathbf{t}_{11}$	$\mathbf{t}_{12}$	Block replication
1	5	1								1	2	1	1
2	3		2		1	1		2	1				1
3	3	1			1		2	3					2
4	4		1		1				2	1	1		1
5	3	1	2					1			3		1
6	3	1	1		2			2		1			1
7	3	1	1		1		1	2			1		1
8	4	1	1							1	3		1
9	3		2			2		2			1		1
10	2	1	2		2			3					2
11	3		1		2			2	1	1			1
12	3	1			2			1	1	2			2
13	2	1	1	1	2			3					1

Table 6.10: D-optimal HGLM unrestricted design, under parameter scenario (b). Table shows incidence within blocks of treatments $\mathbf{t}_1$–$\mathbf{t}_{12}$, defined in Table 6.9. For example, the top-left entry of the main part of the table tells us that treatment $\mathbf{t}_1$ occurs 5 times in the first block.

	X_1	X_2	X_3	X_4
$\mathbf{t}_1$	1	1	1	1
$\mathbf{t}_2$	1	1	1	0
$\mathbf{t}_3$	1	1	0	1
$\mathbf{t}_4$	1	1	0	0
$\mathbf{t}_5$	1	0	1	1
$\mathbf{t}_6$	1	0	1	0
$\mathbf{t}_7$	1	0	0	1
$\mathbf{t}_8$	1	0	0	0
$\mathbf{t}_9$	0	1	1	1
$\mathbf{t}_{10}$	0	1	1	0
$\mathbf{t}_{11}$	0	1	0	1
$\mathbf{t}_{12}$	0	1	0	0
$\mathbf{t}_{13}$	0	0	1	1
$\mathbf{t}_{14}$	0	0	1	0
$\mathbf{t}_{15}$	0	0	0	1

Table 6.11: GLMM unrestricted design, computed under parameter scenario (b) with MQL approximation. Table shows factor settings for the treatments used to define the design in Table 6.12.

	Treatment															
Block	$\mathbf{t}_1$	$\mathbf{t}_2$	$\mathbf{t}_3$	$\mathbf{t}_4$	$\mathbf{t}_5$	$\mathbf{t}_6$	$\mathbf{t}_7$	$\mathbf{t}_8$	$\mathbf{t}_9$	$\mathbf{t}_{10}$	$\mathbf{t}_{11}$	$\mathbf{t}_{12}$	$\mathbf{t}_{13}$	$\mathbf{t}_{14}$	$\mathbf{t}_{15}$	Block replication
1	1	1	1				2		1			1	1	2		1
2	1	2				1	2		1		2		1			1
3		3	1		1		1		1		1		1		1	1
4	1			3			2		1	1			2			1
5	2					1	1	1		1	2		1	1		1
6		1	2	2	1				1				2	1		1
7	1	1		1	1		2		1	1	1		1			1
8	2		1			1		1		1	1		1	1	1	1
9	2		1			2					1	2	2			1
10	1	1		1	2				1		1	1		1	1	1
11	1	1	1		1		1		1	1		1	1		1	1
12	2					3			1		1	1			2	1
13	1			1	2		1		1	1	1	1		1		1
14	1		1	1	1		1		1	1		1	1	1		1
15	1		1		1		2		1	3					1	1
16	2					3		1			3		1			1

Table 6.12: GLMM unrestricted design, computed under parameter scenario (b) with MQL. Table shows incidence of treatments $\mathbf{t}_1$–$\mathbf{t}_{15}$ within blocks. Note that the factor settings for the treatments are given in Table 6.11.

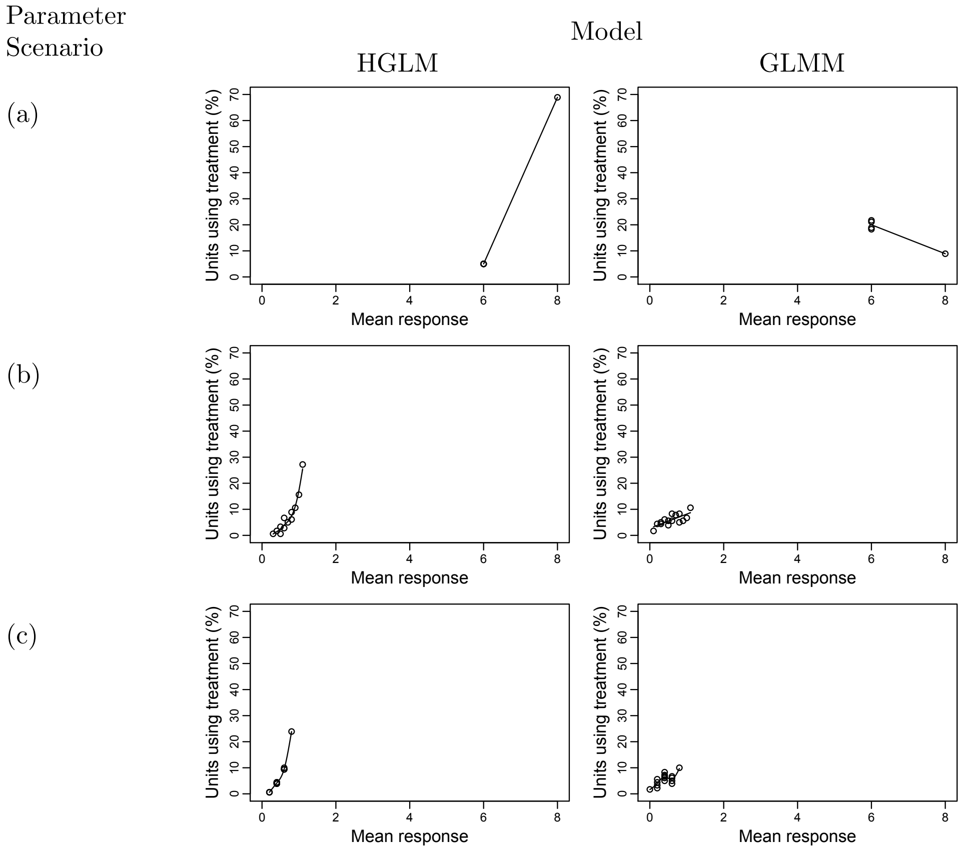

Figure 6.2: Relationship between mean response and replication, unrestricted locally D-optimal designs. Left and right columns correspond to the two different models, HGLM and GLMM (with MQL) respectively. The rows correspond to the different parameter scenarios outlined in the text. Scenario (d) is omitted as any treatment would have the same mean for these values of the parameters.

In the whole-plot designs, many blocks use the treatment which sets all of the factors to the high level, for instance see Table 6.13 which shows the optimal HGLM wholeplot design under parameter scenario (b). This design favours the 'all-high' treatment.

For split-plot designs, the most obvious pattern in the HGLM D-optimal designs is that the whole-plot factors tend to be imbalanced. The favoured value of the whole-plot factor is the one which leads to higher variance. Tables 6.14 and 6.15 show respectively the treatments and their incidence in blocks for the HGLM optimal split-plot design under parameter scenario (a). In Table 6.15, the treatments are ordered so that they can be divided according to whether $x_1 = 1$ or 0. From this it is clear that $x_1 = 0$ in only one of the 16 blocks. This imbalance is not as evident in the GLMM and quasi-likelihood split-plot designs. See for instance Tables 6.16 and 6.17, which together give the quasi-likelihood design under parameter scenario (b).

Replication of high-mean treatments is beneficial in different ways depending on the design structure. Simulation studies were performed to assess the variances and correlations of parameter estimators using the designs from parameter scenario (a). Estimation of both HGLMs and GLMMs was considered, using the `R` packages `hglm` (Ronnegard, Shen and Moudud, 2010) and `glmmML` (Broström, 2011). Using the results of these simulations, the validity of the asymptotic approximations was also evaluated, and found to be satisfactory. For details, see Appendix 6.9.

When estimating the HGLM using an unrestricted design under parameter scenario (b), surprisingly overall the HGLM design has higher estimator variances than the GLMM design. This can be seen in Figure 6.3 which compares diagonal elements of the variance matrix of $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{v}} - \mathbf{v})$ from the two designs. The HGLM design does however have lower correlations between the main effects estimators (Figure 6.4). When using a wholeplot structure, the GLMM design gives much less precise estimation of the random effects in the HGLM (though estimation is better for the other parameters, see Figure 6.5 which again compares the diagonal of the variance matrix for the two designs).

Again under scenario (b), when estimating the GLMM using an unrestricted design structure, the GLMM design provided better (i.e. lower variance) estimators of the fixed effects (Table 6.18). When using wholeplot designs, however, there were convergence issues with the estimation procedure. This seemed to be due to difficulty with the GLMM design in estimating the block effect variance σ^2. This is likely due to the fact that the GLMM design methodology does not yet take into account estimation of this parameter. With σ^2 held fixed at its 'true' value (i.e. $\sigma^2 = 0.1$) during the estimation, the GLMM design provided better estimates of all the parameters, most notably the intercept.

Comparison with HGLM-D_S and quasi-likelihood GLMM designs

Overall, the HGLM-D_S and quasi-likelihood designs were close to D-optimal for estimating the GLMM. Also, HGLM-D_S and quasi-likelihood designs both had similar efficiencies to the GLMM designs for estimating the full HGLM (Tables 6.19–6.22). An apparent exception to the latter is in the results for the split-plot designs, where the GLMM design is somewhat more efficient than the quasi-likelihood for estimating the HGLM. However, this is is not really a proper exception: Table 6.22 shows that in fact the GLMM design was only a 'local solution' of the optimisation problem, being slightly worse for estimating the GLMM than the quasi-likelihood design. It is of course a possibility that there are many designs which are D-optimal for the GLMM which have different D-efficiencies for estimating the HGLM.

Comparison of design structures

Suppose that it is possible to choose between several design structures, but more restricted designs (e.g. split-plot or wholeplot) are economically more convenient. In other words, some of the factors in the experiment are hard, but not impossible, to change within blocks. Then the choice of design will involve a trade-off between the loss in efficiency incurred by using a more restricted design versus the reduction in cost. As a result it is worthwhile quantifying such losses.

Table 6.23 shows the D-efficiencies, for estimation of the HGLM, of the locally D-optimal HGLM designs calculated under parameter scenarios (b), (c) and (d). The use of a split-plot design when we could actually use an unrestricted design results in a loss in D-efficiency of between 14 and 43 percentage points. These losses occur in parameter scenarios (c) and (b) respectively. Clearly the D-efficiency loss in scenario (c) may be tolerable, depending on the reduction in cost achieved. The use of a wholeplot design when we could in fact use an unrestricted design results in larger losses of between 43 and 62 percentage points. Such large losses are less likely to be permissible.

The D-efficiency of a wholeplot design compared to a split-plot design can also be calculated from Table 6.23. For instance, in parameter scenario (c), this efficiency is $57.4/86.0 \times 100\% \approx 66.7\%$. The worst case efficiency loss when using a wholeplot rather than a split-plot design among these examples was 38.8 percentage points. This occurred in scenario (b).

Block	x_1	x_2	x_3	x_4	Repl.
1	1	1	1	1	5
2	1	1	1	0	1
3	1	1	0	1	1
4	1	0	1	1	2
5	1	0	0	1	1
6	0	1	1	1	3
7	0	1	1	0	1
8	0	1	0	1	1
9	0	0	1	1	1

Table 6.13: D-optimal HGLM wholeplot design, parameter scenario (b). Note that the factor settings are held constant across each block, and so only one row per block is required.

	x_1	x_2	x_3	x_4
$\mathbf{t}_1$	1	1	1	1
$\mathbf{t}_2$	1	1	1	0
$\mathbf{t}_3$	1	1	0	1
$\mathbf{t}_4$	1	1	0	0
$\mathbf{t}_5$	1	0	1	1
$\mathbf{t}_6$	1	0	1	0
$\mathbf{t}_7$	1	0	0	1
$\mathbf{t}_8$	1	0	0	0
$\mathbf{t}_9$	0	1	1	1
$\mathbf{t}_{10}$	0	1	1	0
$\mathbf{t}_{11}$	0	1	0	0
$\mathbf{t}_{12}$	0	0	1	1
$\mathbf{t}_{13}$	0	0	0	1
$\mathbf{t}_{14}$	0	0	0	0

Table 6.14: D-optimal HGLM splitplot design, parameter scenario (a). Table shows factor settings for the treatments used in the design defined by Table 6.15.

	Treatment														
	$x_1 = 1$								$x_1 = 0$						
Block	$\mathbf{t}_1$	$\mathbf{t}_2$	$\mathbf{t}_3$	$\mathbf{t}_4$	$\mathbf{t}_5$	$\mathbf{t}_6$	$\mathbf{t}_7$	$\mathbf{t}_8$	$\mathbf{t}_9$	$\mathbf{t}_{10}$	$\mathbf{t}_{11}$	$\mathbf{t}_{12}$	$\mathbf{t}_{13}$	$\mathbf{t}_{14}$	Block replication
1	3		2	1	2		1	1							1
2	4	3	1		1		1								1
3	2	1			3	2		2							1
4	3		2		3	1		1							1
5	1	2	2		1	2	2								1
6	2		2	2	1	2	1								1
7	3	4	1				1	1							1
8									2	1	1	1	2	3	1
9	1	1	1		4	2	1								1
10	2		2	1		3	2								1
11	2	1	3	1	1	1	1								1
12	1	3		1	3		2								1
13	1	4		2	1		2								1
14	3		3		2	1	1								1
15	3		1	1	2		3								1
16	5	2	1	1			1								1

Table 6.15: D-optimal HGLM splitplot design, parameter scenario (a). Table shows incidence of treatments $\mathbf{t}_1$–$\mathbf{t}_{14}$ within blocks. Note that the factor settings for the treatments are given in Table 6.14.

	x_1	x_2	x_3	x_4
$\mathbf{t}_1$	1	1	1	1
$\mathbf{t}_2$	1	1	1	0
$\mathbf{t}_3$	1	1	0	1
$\mathbf{t}_4$	1	1	0	0
$\mathbf{t}_5$	1	0	1	1
$\mathbf{t}_6$	1	0	1	0
$\mathbf{t}_7$	1	0	0	1
$\mathbf{t}_8$	1	0	0	0
$\mathbf{t}_9$	0	1	1	1
$\mathbf{t}_{10}$	0	1	1	0
$\mathbf{t}_{11}$	0	1	0	1
$\mathbf{t}_{12}$	0	1	0	0
$\mathbf{t}_{13}$	0	0	1	1
$\mathbf{t}_{14}$	0	0	1	0
$\mathbf{t}_{15}$	0	0	0	1
$\mathbf{t}_{16}$	0	0	0	0

Table 6.16: Quasi-likelihood GLMM split-plot design, parameter scenario (b). Table shows factor settings for the treatments used in the design defined in Table 6.17.

	Treatment																
	$x_1 = 1$								$x_1 = 0$								
Blo-ck	$\mathbf{t}_1$	$\mathbf{t}_2$	$\mathbf{t}_3$	$\mathbf{t}_4$	$\mathbf{t}_5$	$\mathbf{t}_6$	$\mathbf{t}_7$	$\mathbf{t}_8$	$\mathbf{t}_9$	$\mathbf{t}_{10}$	$\mathbf{t}_{11}$	$\mathbf{t}_{12}$	$\mathbf{t}_{13}$	$\mathbf{t}_{14}$	$\mathbf{t}_{15}$	$\mathbf{t}_{16}$	Rep.
1	1	1	2	1	2	1	1	1									1
2	1	1	1	1	1	2	2	1									1
3	2		2	1	1	1		3									1
4									2	1	1	1	1	2	1	1	1
5	1	1	3	1	2	1		1									1
6									1	2	2	1	2		2		1
7		3	1	2	2	2											1
8									3	1	1	1	1	1	2		1
9									1		2	1		2	2	2	1
10									3	2				1	3	1	1
11									1			2	3	1	1	2	1
12	2	1	2		1	2	1	1									1
13	2	2	1	2	2		1										1
14	1	1	2		2	1	1	2									1
15									2	2	1		1	2	1	1	1
16	3	2		1		1	3										1

Table 6.17: Quasi-likelihood GLMM split-plot design, parameter scenario (b). Table shows incidence of treatments $\mathbf{t}_1$–$\mathbf{t}_{16}$ within blocks. Note that the factor settings for the treatments are given in Table 6.16.

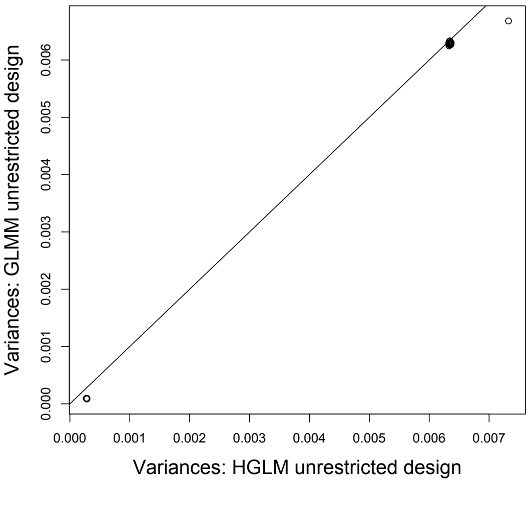

Figure 6.3: Variances of HGLM parameter estimators, comparison between unrestricted D-optimal HGLM and GLMM designs in parameter scenario (b). Each point on the plot corresponds to an entry in the diagonal of the covariance matrix, $\text{var}(\boldsymbol{\beta}, \hat{\mathbf{v}} - \mathbf{v})$.

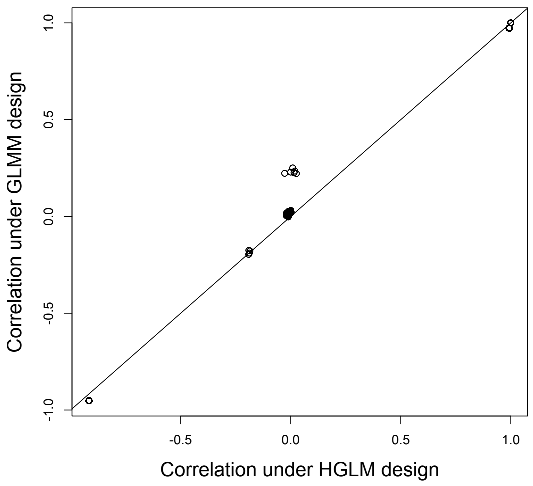

Figure 6.4: Correlations between HGLM parameter estimators, comparison between unrestricted D-optimal HGLM and GLMM designs in parameter scenario (b). Each point on the plot corresponds to an off-diagonal term in the correlation matrix, $\text{corr}(\boldsymbol{\beta}, \hat{\mathbf{v}} - \mathbf{v})$.

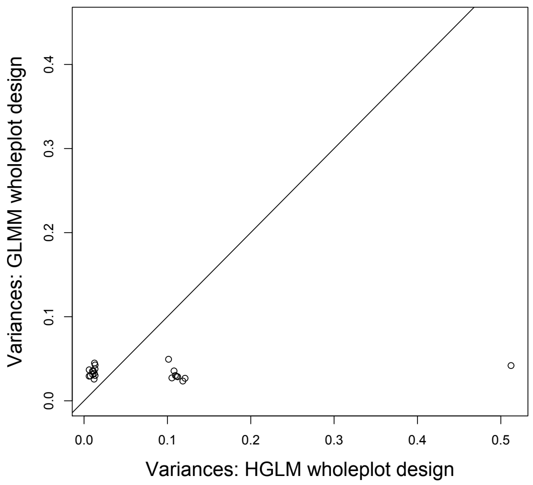

Figure 6.5: Jittered variances for estimating HGLM, comparison between wholeplot D-optimal HGLM and GLMM designs in parameter scenario (b). Each point on the plot corresponds to an entry in the diagonal of the covariance matrix, $\text{var}(\boldsymbol{\beta}, \hat{\mathbf{v}} - \mathbf{v})$

	Design	
Parameter	HGLM-D	GLMM
β_0	0.5085	0.0447
β_1	0.1144	0.0274
β_2	0.1180	0.0270
β_3	0.1153	0.0267
β_4	0.1144	0.0282

Table 6.18: Maximum likelihood estimator variances, for estimation of GLMM, using unrestricted designs under parameter scenario (b).

	Design	
Scenario ($\boldsymbol{\beta}$-values)	GLMM	HGLM-D$_S$
(b)	93.6	93.6
(c)	96.5	98.7
(d)	100.0	100.0

Table 6.19: D-efficiency (%), for estimation of the HGLM, of unrestricted designs

	Design	
Scenario ($\boldsymbol{\beta}$-values)	HGLM-D	HGLM-D$_S$
(b)	83.4	100.0
(c)	88.7	100.0
(d)	100.0	100.0

Table 6.20: MQL D-efficiency (%), for estimation of the GLMM, of unrestricted designs

	Design	
Scenario ($\boldsymbol{\beta}$-values)	QL	GLMM
(b)	92.7	98.7
(c)	92.0	93.9
(d)	99.9	99.9

Table 6.21: D-efficiency (%), for estimation of the HGLM, of split-plot designs

	Design	
Scenario ($\boldsymbol{\beta}$-values)	QL	HGLM
(b)	100.4	95.5
(c)	99.5	94.5
(d)	99.9	100.0

Table 6.22: MQL D-efficiency (%), for estimation of the GLMM, of split-plot designs

	Design structure	
Scenario ($\boldsymbol{\beta}$-values)	Split-plot	Wholeplot
(b)	57.5	38.0
(c)	72.7	44.5
(d)	86.0	57.4

Table 6.23: D-efficiency (%), for estimation of HGLM, of locally D-optimal restricted HGLM designs compared with unrestricted locally D-optimal HGLM design under parameter scenarios (b), (c) and (d).

6.6.3 Bayesian design

In this section, we consider split-plot designs for the first order model (6.7) which are robust to a range of possible parameter values. We assume that x_1 is the only whole-plot factor. To attempt to find a more robust solution, we find the design maximising the objective function (6.6). As regards prior beliefs, we adopt independent uniform prior distributions with ranges $[-0.5, 0.5]$, $[-0.3, 0.3]$, $[-0.2, 0.2]$ and $[-0.1, 0.1]$ for β_1, β_2, β_3 and β_4 respectively. It is assumed a priori that $\beta_0 = 0$ and $\alpha_1 = 10$. This corresponds to a moderate range of uncertainty about the possible size and direction of the effects of the factors, with the whole-plot factor potentially having a non-negligible influence. To quantify this, suppose that $\beta_1 = 0.5$. If x_1 changes from 0 to 1, then ceteris paribus the conditional mean of the response will be multiplied by $e^{0.5} \approx 1.64$, a 64% increase. For $\beta_1 = -0.5$, the same shift in x_1 would result in a 39.3% decrease in the conditional mean response.

We approximated the integral in objective function (6.6) by

$$\psi_{\text{Bayes}}(\xi) \approx \frac{1}{30} \sum_{s=1}^{30} \log \psi_D(\xi; \boldsymbol{\beta}_s, \boldsymbol{\alpha}) \,,$$

where $\boldsymbol{\beta_s}$, $s = 1, \ldots, 30$, is a 30-point Latin Hypercube sample from

$$[-0.5, 0.5] \times [-0.3, 0.3] \times [-0.2, 0.2] \times [-0.1, 0.1] \,.$$

This number of abscissae rendered the optimisation feasible, but still a substantial task requiring several hours of computation. Quadrature schemes involving a larger number of abscissae, such as that of Gotwalt et al. (2009) do not currently seem feasible with designs containing this many free co-ordinates.

The local efficiency, at $\boldsymbol{\beta} = \mathbf{0}$, of the Bayesian design was 99.99%, so there is very little difference between the Bayesian and centroid designs. Figure 6.6 compares the Bayesian design with the locally optimal design at this centroid in terms of the efficiency distributions induced by the prior distribution. To produce this figure we first sampled 450 parameter vectors from the prior distribution, computing the locally optimal design at each of these parameter vectors. The efficiency of the Bayesian design and centroid design were then computed by comparing to the locally optimal design. Finally, kernel density estimates of the two efficiency distributions were made. From the figure, we see that the performance of the Bayesian design is essentially indistinguishable from the centroid design in this case. Both designs are very robust, suggesting that in practical terms the degree of parameter uncertainty expressed in the prior is not hugely significant.

6.7 Discussion

Optimal HGLM designs are a feasible strategy in problems where there are blocks or split-plot structures and count responses, although finding unrestricted Bayesian designs for the 7-factor example is a computational challenge. They provide an interesting alternative compared with GLMM designs along the lines of Chapter 2. Efficiency gains are clearly possible when using this approach in favour of more naïve ones, particularly through the replication of high-variance

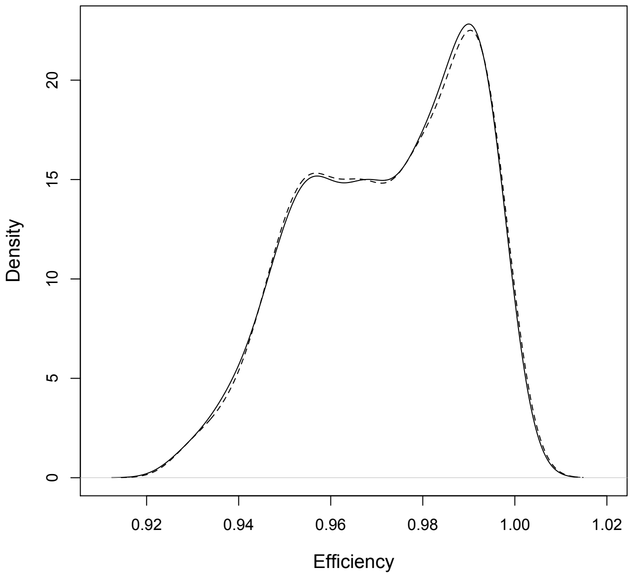

Figure 6.6: Kernel density estimates of the efficiency distributions of the Bayesian design (solid line) and locally optimal design at the centroid (dotted line) from Section 6.6.3.

treatments.

An open issue, as with the GLMM design scenario, is the consideration of the quality of the estimation of variance components. Lee and Nelder (1996, Section 4.1) comment that estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are asymptotically orthogonal, and nearly so in finite samples. However this may not hold when the blocks are very small, and in this case it is likely to be difficult to estimate $\boldsymbol{\alpha}$.

In our examples, we have focussed on the Poisson-gamma model, since this yields analytically tractable expressions for the matrices $\mathrm{E}(W)$ and $\mathrm{E}(U)$ which are present in the marginal h-information matrix. Other models will typically require additional computation in order to approximate these matrices. Investigation of methods and examples for other response and random effects distributions is an avenue for future research.

6.8 Appendix: Further details

In this section we give explicit analytical expressions for the terms $\mathrm{E}(W)$ and $\mathrm{E}(U)$ which appear in the marginal h-information matrix. For other HGLMs these expectations can typically not be evaluated analytically, and some approximation procedure is required.

We have that $y_{ij}|u_i \sim \text{Poisson}(\mu_{ij})$ where $\mu_{ij} = \exp\{\mathbf{f}^T(\mathbf{x}_{ij})\boldsymbol{\beta}\}u_i$ and $u_i \sim \text{Gamma}(\alpha_1, \alpha_2)$. Here $\alpha_1 > 0$ is a shape parameter, and $\alpha_2 > 0$ is a scale parameter. As the model uses the canonical link, the GLM weight (6.4) reduces to

$$V(\mu_{ij}) = \mu_{ij} = \exp\{\mathbf{f}^T(\mathbf{x}_{ij})\}u_i \,.$$

Since $\mathrm{E}(u_i) = \alpha_1\alpha_2$, which is equal to 1 by the restriction suggested in Lee and Nelder (1996) to ensure identifiability,

$$\mathrm{E}(W) = \operatorname{diag}\left\{\exp[\mathbf{f}^T(\mathbf{x}_{ij})] : (i,j) \text{ in lexicographical order}\right\} .$$

The matrix U is diagonal with ith entry $-\partial^2 \log f_{\mathbf{v}}(\mathbf{v};\boldsymbol{\alpha})/\partial v_i^2$. The pdf for the random effect u_i is

$$f_u(u_i;\alpha_1,\alpha_2) = \frac{u_i^{\alpha_1-1}\exp(-u_i/\alpha_2)}{\Gamma(\alpha_1)\alpha_2^{\alpha_1}} \,.$$

Using the fact that $v_i = \log u_i$, and applying the transformation rule for pdfs, we see that the density of v satisfies

$$\begin{aligned}\log f_v(v) &= \log f_u(\exp(v)) + v \\ &= [(\alpha_1 - 1)\log u - u/\alpha_2 - \log(\Gamma(\alpha_1)\alpha_2^{\alpha_1})] + v \\ &= \alpha_1 v - \exp(v)/\alpha_2 - \log(\Gamma(\alpha_1)\alpha_2^{\alpha_1}) \,.\end{aligned}$$

Hence

$$\begin{aligned}\mathrm{E}\left(-\frac{\partial^2 \log f_{\mathbf{v}}(\mathbf{v};\boldsymbol{\alpha})}{\partial v_i^2}\right) &= \mathrm{E}\left(-\frac{\partial^2}{\partial v_i^2}\sum_{i'=1}^{n_b}\log f_v(v_{i'};\boldsymbol{\alpha})\right) \\ &= \mathrm{E}\left(\exp(v_i)/\alpha_2\right) \\ &= \alpha_1 \,,\end{aligned}$$

where the first line follows by independence of the random effects. Thus $\mathrm{E}(U)$ is in fact an $n_b \times n_b$ diagonal matrix with all entries equal to α_1.

6.9 Appendix: Validation of asymptotic approximations

The simulation studies conducted in Section 6.6.2 were also used to assess the validity of the asymptotic approximations. The sampled parameter estimates were used to compute a Monte Carlo approximation, V_{emp}, of $\operatorname{var}(\hat{\boldsymbol{\beta}}^T, \hat{\mathbf{v}}^T - \mathbf{v}^T)^T$. In Figures 6.7 and 6.8, the entries of V_{emp} are plotted against the corresponding entries in the theoretical approximation, V_{th}, to $\operatorname{var}(\hat{\boldsymbol{\beta}}^T, \hat{\mathbf{v}}^T - \mathbf{v}^T)^T$. The latter is obtained by inverting the marginal h-information matrix, (6.5), in other words $V_{\text{th}} = J_M^{-1}$. The plots shown are for the unrestricted and whole-plot designs under parameter scenario (a). Other examples with comparable numbers and sizes of blocks resulted in similar figures. The points in the figures are close to the line of equality, indicating that the asymptotic approximations hold reasonably well, in other words $V_{\text{emp}} \approx V_{\text{th}}$.

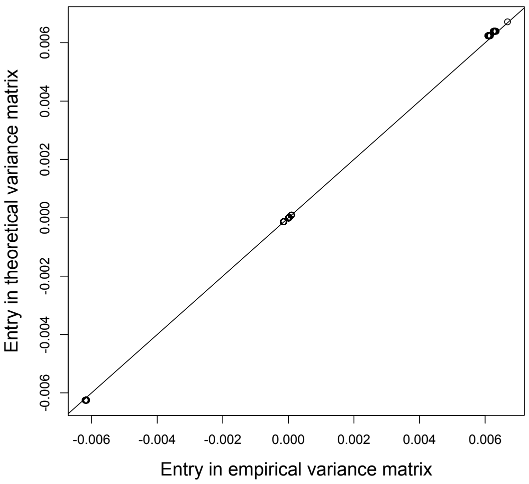

Figure 6.7: Comparison between theoretical and empirical covariance matrices, HGLM unrestricted design under parameter scenario (a).

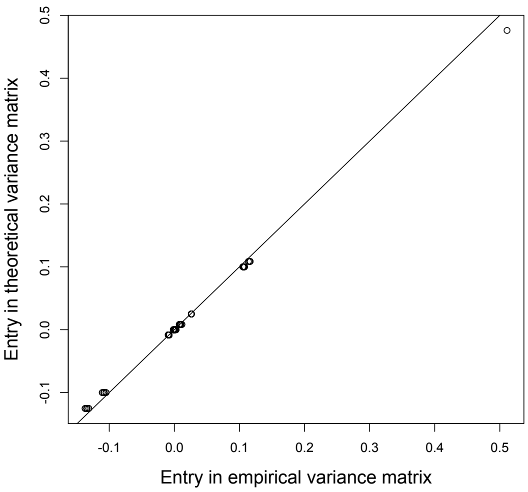

Figure 6.8: Comparison between theoretical and empirical covariance matrices, HGLM wholeplot design under parameter scenario (a).

Chapter 7

Optimal design in the vicinity of singularities

In this chapter we consider in more depth the construction of designs which are robust to parameter uncertainty. In particular, we show that the most popular formulation of the Bayesian D-optimality criterion (Chaloner and Larntz, 1989; Chaloner and Verdinelli, 1995) is degenerate in a wider range of scenarios than has previously been explicitly acknowledged. To overcome this problem, we consider (i) the use of alternative optimality criteria which are better behaved, and (ii) the use of designs with infinite support, defined through a probability density function.

7.1 Introduction

In recent years there has been focus on applying more complex statistical models to the analysis of experimental data. For instance, in some industrial applications the response variable does not follow a normal distribution, and so a generalised linear model (GLM) or generalised linear mixed model (GLMM) is appropriate for the analysis. Examples in the photography, semiconductor and aeronautrics industries are given by Robinson et al. (2004, 2006) and Woods and Van de Ven (2011). Another area is pharmacokinetics, where it is beneficial to use mechanistic models whose parameters have a direct biological interpretation. Typically compartmental models are used, with random effects which model the variation in drug response between patients (see Retout, Comets, Samson and Mentré, 2007, for references and a discussion of optimal design for these models).

These more complex models have in common the property that the D-optimal design may depend on the unknown values of the model parameters, $\boldsymbol{\theta} \in \mathbb{R}^p$. Several methods have been proposed to derive designs which are reasonably efficient under a range of plausible values for the parameters, such as maximin and Bayesian approaches. We initially consider approximate designs defined by a discrete probability measure. For a univariate response we can write

$$\xi = \left\{ \begin{array}{ccc} \mathbf{x}_1 & \dots & \mathbf{x}_k \\ w_1 & \dots & w_k \end{array} \right\}, \tag{7.1}$$

with $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^q$ and $w_i > 0$ for $1 \leq i \leq k$, and $\sum_{i=1}^k w_i = 1$. If there are blocks present within the data, or there are multivariate responses, we may need a slightly different notion of design, such as that in Chapters 2 and 3. The weights w_i represent the proportions of units which are assigned to experimental conditions $\mathbf{x}_i$.

A definitive account of Bayesian design is given by Chaloner and Verdinelli (1995), who argue that a principled approach is to seek the design, ξ^*, which maximises the expected gain in Shannon information from prior to posterior density, given by

$$\phi_S(\xi) = \int \log \frac{p(\boldsymbol{\theta}|\mathbf{y},\xi)}{f(\boldsymbol{\theta})}\, p(\mathbf{y},\boldsymbol{\theta}|\xi)\, d\boldsymbol{\theta}\, d\mathbf{y}\,, \tag{7.2}$$

where $f(\boldsymbol{\theta})$ is the prior density for $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\mathbf{y},\xi)$ is the posterior density of $\boldsymbol{\theta}$, and $p(\mathbf{y},\boldsymbol{\theta}|\xi)$ is the joint density of $\mathbf{y}$ and $\boldsymbol{\theta}$. For nonlinear models, the integral (7.2) is intractable. In this case, the authors propose maximising one of the following two substitutes, based on asymptotic normal approximations to the posterior distribution,

$$\phi(\xi) = E_{\boldsymbol{\theta}} \log |nM(\xi;\boldsymbol{\theta})| \tag{7.3}$$

$$\phi_2(\xi) = E_{\boldsymbol{\theta}} \log |nM(\xi;\boldsymbol{\theta}) + R|\,, \tag{7.4}$$

where M is the Fisher information matrix, n is the sample size, and R is the matrix of second derivatives of $\log f$ or the prior precision matrix. The resulting criteria are referred to as Bayesian D-optimality criteria.

The objective function (7.3) is also often used when there is not necessarily an assumption that the resulting data will be analysed using Bayesian methods (e.g. Woods et al., 2006), or when different priors may be used for the design and analysis. In either of these cases, the principled justification of the use of (7.2) breaks down somewhat. Instead, we can think of maximisation of (7.3) essentially as trading off the efficiency of the design across the likely values of the parameters.

We say that $\boldsymbol{\theta}_0$ is a *singularity* if, for any fixed ξ, $|M(\xi;\boldsymbol{\theta})| \to 0$ as $\boldsymbol{\theta} \to \boldsymbol{\theta}_0$. If it is possible a priori for $\boldsymbol{\theta}$ to be arbitrarily close to $\boldsymbol{\theta}_0$, then there may be serious issues with the use of (7.3), or also (7.4) if R is not positive definite. Specifically, we may have that $\phi(\xi) = -\infty$, in other words (7.3) does not converge, for many designs which would not traditionally be considered to be singular. In extreme cases, see for instance Section 7.6, we can have that $\phi(\xi) = -\infty$ for all finitely supported designs, even though almost all such designs have positive (local) efficiency for all values of $\boldsymbol{\theta}$ which are possible a priori. In this situation, (7.3) fails to discriminate between any proposed design, and so is useless in helping us make a choice.

This issue of convergence has been little discussed in the optimal design literature, and yet it is important to understand in order to be able to design experiments when there is 'extreme' parameter uncertainty. There is an example of non-convergence in Tsutakawa (1972), and a small amount of discussion of the issue in Chaloner and Verdinelli (1995) who mention it only when the prior support is unbounded; in Section 7.6 we present an example of non-convergence when the prior is supported on a bounded interval. In this chapter, we attempt to outline the main problems surrounding convergence and find examples from the literature in which the issue may arise. We also put forward some suggestions for alternative methods of handling parameter uncertainty which work in the vicinity of singularities. Several analytical results are given for the exponential model, which serves to help understanding of the general issues.

One potential way forward is to extend the traditional notion of an approximate design. Rather than restricting our attention only to finitely supported probability measures as in Atkinson et al. (2007), we also consider designs with infinite support defined by a probability density function. We refer to such designs as density designs, and the performance of finite samples from these densities is considered.

7.2 Singularities

We now attempt to clarify, at least conceptually, the issue of convergence. Let ξ be a fixed design and $\boldsymbol{\theta}_0$ be a singularity such that there are values of $\boldsymbol{\theta}$ which are arbitrarily close to $\boldsymbol{\theta}_0$, and which are possible a priori. Then as $\boldsymbol{\theta} \to \boldsymbol{\theta}_0$, $\log|M(\xi;\boldsymbol{\theta})| \to -\infty$. Depending on the rate of the convergence of $\log|M|$ to $-\infty$, and the behaviour of f near $\boldsymbol{\theta}_0$, the integral in (7.3) may or may not converge.

For instance, suppose that the model had one parameter, $p = 1$, $\theta_0 = 0$ and $\log|M|$ and f were such that

$$\log|M(\xi,\theta)|f(\theta) = O(|\theta|^{-1}) \quad \text{as } \theta \to 0\,.$$

Then we could approximate the integral (7.3) close to $\theta_0 = 0$, as

$$\begin{aligned}\int_{-\delta}^{\delta} \log|M(\xi,\theta)|f(\theta)d\theta &\approx C\int_{-\delta}^{\delta} |\theta|^{-1}d\theta \\ &= -\infty\,,\end{aligned}$$

or some C, which is negative since $\log|M(\xi;\theta)| \to -\infty$. Since in regular problems $\log|M|$ is bounded above, this is sufficient to establish that $\phi(\xi) = -\infty$. The approximation can be made more rigorous, but the above sketch is enough to show what might happen.

Essentially, for ϕ to be finite, the prior density must decay sufficiently quickly in the neighbourhood of all singularities. It is difficult in general to determine analytically what is the necessary rate of decay for a given model. If we do not consider the issue, we run the risk that we are in the situation discussed in Section 7.6, in which all finitely supported designs have $\phi(\xi) = -\infty$.

A problem in practice is that we cannot establish numerically whether the integral (7.3) converges. We may perform checks, for instance, by using quadrature schemes with an increasing number, n_a, of abscissae – but we will be unable to tell if it is the case that our estimate of $\phi(\xi)$ is simply converging very slowly to $-\infty$ as n_a increases.

7.2.1 Examples where the issue arises

In this section we give examples of some models which contain singularities, and one which may potentially do so. Further details are given in Sections 7.3 and 7.6.

Logistic model

Chaloner and Larntz (1989) consider Bayesian designs for the logistic model where the response is Bernoulli, with a single explanatory variable x, parameters $\boldsymbol{\theta}^T = (\mu, \beta)$ and probability of success

$$p(x, \boldsymbol{\theta}) = \frac{1}{1 + \exp\{-\beta(x - \mu)\}} \,. \tag{7.5}$$

In this model there are singularities at $\beta = 0, \infty$. From here onwards, we use the phrase '$\beta = \infty$ is a singularity' as a shorthand for: as $\beta \to \infty$, for any fixed design $|M(\xi, \boldsymbol{\theta})| \to 0$. The reason for the occurrence of the singularity at 0 is simple, since when $\beta = 0$, μ is not identifiable. For further details, see Section 7.3.1.

Binary GLMMs

Let us suppose that we have binary responses in blocks, as occurs in the aeronautical industry example of Woods and Van de Ven (2011). Such data can be modelled using a binary GLMM. Let the jth response in the ith block be denoted by y_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m_i$, and corresponding vectors of explanatory variables by $\mathbf{x}_{ij} \in [-1, 1]^q$. Then the random intercept binary generalised linear mixed model is given by

$$\begin{aligned} y_{ij}|u_i &\sim \text{Bernoulli}\{\mu(\mathbf{x}_{ij}|u_i)\} \\ g(\mu(\mathbf{x}|u)) &= \mathbf{f}^T(\mathbf{x})\boldsymbol{\beta} + u \\ u_i &\sim N(0, \sigma^2)\,, \end{aligned} \tag{7.6}$$

where g is the logit function, and the u_i are random intercepts which are independent for different i. The (known) function $\mathbf{f} : [-1, 1]^q \to \mathbb{R}^p$ maps the explanatory variables to the terms in the model, and $\boldsymbol{\beta}$ is the vector of p fixed effects parameters. Design for these models is considered in Chapters 2 and 3. For large values of σ^2, the model is degenerate (see Section 7.9.4). This causes concern for the potential existence of a singularity at $\sigma^2 = \infty$. Moreover, it is a possibility that similar problems to the 1-factor logistic model may occur at $\boldsymbol{\beta} = \infty$.

Exponential model

The exponential decay model occurs in chemical kinetics (Atkinson et al., 2007, pp. 248–250). We consider the model parameterised by half-life, $\theta \in \mathbb{R}$, rather than the 'rate' parameterisation used by the above authors. The response y is the concentration of a chemical compound, and the explanatory variable is time, denoted here by $x \geq 0$. The model is given by

$$\begin{aligned} &y_i = \eta(x_i, \theta) + \epsilon_i \qquad \eta(x, \theta) = e^{-x/\theta} \\ &\epsilon_i \sim N(0, \sigma^2)\,, \end{aligned} \tag{7.7}$$

where $1 \leq i \leq n$, $x_i \geq 0$, and $\sigma^2 > 0$. The model has singularities at $\theta = 0, \infty$.

Compartmental model

Atkinson, Chaloner, Herzberg and Juritz (1993) consider designs for a single compartmental model, which can be used to model the passage of a drug through a subject. The response y

is the concentration in the blood of a chemical of interest, and $x \geq 0$ is time. The model has parameters $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$ and is given by

$$\begin{aligned} y_i &= \eta(x_i, \boldsymbol{\theta}) + \epsilon_i \\ \eta(x, \theta) &= e^{-\theta_1 x} - e^{-\theta_2 x} \\ \epsilon_i &\sim N(0, \sigma^2)\,, \end{aligned} \tag{7.8}$$

where $1 \leq i \leq n$, $x_i \geq 0$, and $\sigma^2 > 0$. We have omitted the magnitude parameter θ_3 from the model as this does not affect the optimal design. This model has singularities along the line $\theta_1 = \theta_2$ and when $\theta_1 = 0$ or $\theta_2 = 0$.

7.3 Examples in more detail

In this section we justify our claims about the existence of singularities in two of the examples, in varying levels of mathematical rigour.

7.3.1 Logistic model

Chaloner and Larntz (1989) give the expression for the determinant of the information matrix of model (7.5) for a design

$$\xi = \left\{ \begin{array}{ccc} x_1 & \cdots & x_k \\ w_1 & \cdots & w_k \end{array} \right\}. \tag{7.9}$$

as

$$|M(\xi, \boldsymbol{\theta})| = \beta^2 ts\,, \tag{7.10}$$

where

$$\begin{aligned} t &= \sum_{i=1}^{k} w_i \lambda_i & s &= \sum_{i=1}^{k} w_i \lambda_i (x_i - \bar{x})^2 \\ \bar{x} &= t^{-1} \sum_{i=1}^{k} w_i \lambda_i x_i & \lambda_i &= p(x_i, \boldsymbol{\theta})\{1 - p(x_i, \boldsymbol{\theta})\}\,, \end{aligned}$$

and $p(x, \boldsymbol{\theta})$ is given by (7.5).

To see that there is a singularity at $\beta = 0$, first note that $\min_i x_i \leq \bar{x} \leq \max_i x_i$ and $0 \leq \lambda_i \leq 1/4$ for all i. Therefore

$$\begin{aligned} s &\leq (1/4)(\max_i x_i - \min_i x_i)^2 \\ t &\leq 1/4\,, \end{aligned}$$

and clearly also $s, t \geq 0$. Thus

$$|M(\xi, \boldsymbol{\theta})| \leq \beta^2 (1/16)(\max_i x_i - \min_i x_i)^2 \to 0 \quad \text{as } \beta \to 0\,.$$

The existence of this singularity is due to the fact that at $\beta = 0$, μ is not identifiable.

The singularity as $\beta \to \infty$ occurs because the mean dose-response curve converges to a step function centred on $x = \mu$. Thus any fixed design will eventually miss the region of interest for β sufficiently large, as the 'jump' in the mean response will occur between design points. For the formal proof, see Section 7.9.3.

The existence of the singularity at $\beta = 0$ is due to the choice of parameterisation, which is such that model (7.5) is not a GLM (as the model is not linear in the parameters on the scale of the linear predictor). This issue could be avoided by choosing a GLM parameterisation. However, the scale given is probably the most natural on which to specify prior information about the parameters: if one assumed a priori that the intercept and slope parameters were independent, this would imply that β and μ were correlated (possibly only weakly). This may not necessarily be desirable. In any case, the chosen scale is more than likely the one we would use to report results, and so it is sensible to use this parameterisation for the design.

To consider the impact of the prior distribution on the convergence of (7.3), we fix a particular design

$$\xi = \left\{ \begin{array}{cc} 0.5 & 0.5 \\ -1 & 1 \end{array} \right\} ,$$

and let $\mu = 0$.

For this design the determinant of the information matrix is

$$|M(\xi, \boldsymbol{\theta})| = |\beta|^2 e^{2|\beta|}(1 + e^{|\beta|})^{-4} .$$

Since $0 \leq |\beta| \leq e^{|\beta|}$ and $e^{|\beta|} \geq 1$, the above satisfies

$$|\beta|^4 (2e^{|\beta|})^{-4} \leq |M(\xi, \boldsymbol{\theta})| \leq e^{4|\beta|} ,$$

and so

$$4 \log |\beta| - 4 \log 2 - 4|\beta| \leq \log |M(\xi, \boldsymbol{\theta})| \leq 4|\beta| . \tag{7.11}$$

Suppose that $X_1(\beta)$ and $X_2(\beta)$ are $\mathbb{R}$-valued functions satisfying $X_1(\beta) \leq X_2(\beta)$ for all β. Then we have the following property

$$E(X_1) = \int_{\mathbb{R}} X_1(\beta) f(\beta) d\beta \leq \int_{\mathbb{R}} X_2(\beta) f(\beta) d\beta = E(X_2) . \tag{7.12}$$

Using (7.11) and (7.12) with $X_1 = \log |M|$ and $X_2 = 4|\beta|$, we see that if $E(|\beta|) < \infty$ then also $E \log |M| \leq 4E(|\beta|) < \infty$. The condition $E(|\beta|) < \infty$ is clearly quite unrestrictive and is only broken by very pathological priors, such as the Cauchy distribution.

To ensure also that $E \log |M| > -\infty$, a sufficient extra condition is that $E \log |\beta| > -\infty$. We see this by considering (7.11) and (7.12) with $X_1 = 4 \log |\beta| - 4 \log 2 - 4|\beta|$ and $X_2 = \log |M|$. Similar arguments show that this extra condition is also necessary. However, it is rather more strict than the first condition, and it therefore may be violated more easily by prior distributions which are only 'semi-pathological'. For instance, if $\beta = \exp(-|Z|)$ where Z has a standard Cauchy distribution, then this condition is broken and so $\phi(\xi) = -\infty$. Nonetheless, this distribution on β might be plausible for some applications. Its cumulative distribution function is

$$F(\beta) = 2\left(\frac{-\arctan\{-\log(\beta)\}}{\pi} + 0.5\right), \quad 0 < \beta \le 1\,. \tag{7.13}$$

This CDF is plotted in Figure 7.1.

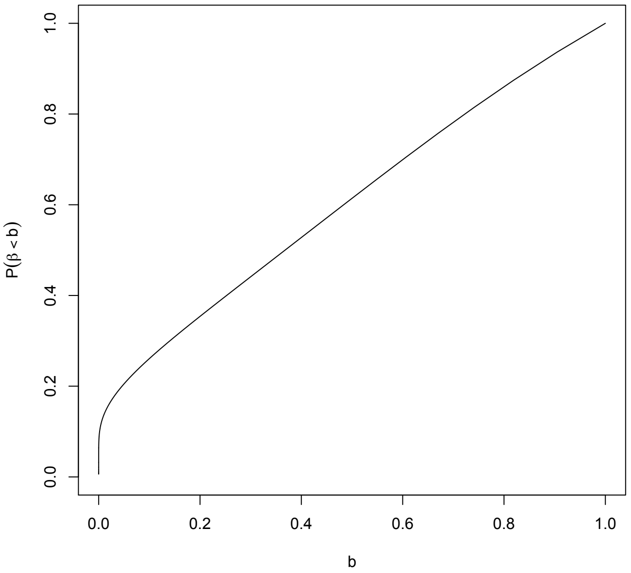

Figure 7.1: Cumulative distribution function, (7.13), of β under the semi-pathological prior in Section 7.3.1

7.3.2 Compartmental model

The information matrix for model (7.8) under design ξ, using the notation in (7.9), is

$$\begin{aligned} M(\xi, \boldsymbol{\theta}) &= \sum_{i=1}^{k} w_i \left(\frac{\partial \eta(x_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \eta(x_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T \\ &= \begin{pmatrix} \sum_{i=1}^k w_i x_i^2 e^{-2\theta_1 x_i} & -\sum_{i=1}^k w_i x_i^2 e^{-(\theta_1+\theta_2)x_i} \\ -\sum_{i=1}^k w_i x_i^2 e^{-(\theta_1+\theta_2)x_i} & \sum_{i=1}^k w_i x_i^2 e^{-2\theta_2 x_i} \end{pmatrix}. \end{aligned}$$

As $\theta_1 \to \theta_2$, the first column tends to the negative of the second column and so $|M(\xi, \boldsymbol{\theta})| \to 0$. Thus there is a singularity along the line $\theta_1 = \theta_2$. This singularity arises because the expected value of the response is identical for any pair of (θ_1, θ_2) such that $\theta_1 = \theta_2$.

If either θ_1 or $\theta_2 \to \infty$, three entries of the information matrix will tend to 0 and so $|M(\xi, \boldsymbol{\theta})| \to 0$. Therefore there are also singularities where $\theta_1 = \infty$ or $\theta_2 = \infty$. These singularities occur because the part of the model corresponding to the large parameter value is eventually indistinguishable from 0 on the (fixed) design points. The prior distributions used by Atkinson et al. (1993), which were uniform priors, successfully avoided all the singularities we have identified, but there was no mention of the issue.

7.4 Potential solutions

In this section we outline our main ideas for obtaining designs in the case where there are singularities, and so the issue of convergence of (7.3) is unavoidable. However to summarise, we must either give up the primacy of the criterion that maximises (7.3), or else in general be prepared to abandon finitely supported designs.

7.4.1 Local efficiency distribution

Our first suggestion, which originated in Woods et al. (2006), is that the assessment of any candidate design ξ should be on the basis of the local efficiency function $\mathrm{eff}(\xi|\boldsymbol{\theta})$, and the distribution of local efficiencies induced by our prior beliefs on $\boldsymbol{\theta}$. The local efficiency is given by

$$\mathrm{eff}(\xi|\boldsymbol{\theta}) = \frac{|M(\xi,\boldsymbol{\theta})|^{1/p}}{\sup_{\xi'} |M(\xi',\boldsymbol{\theta})|^{1/p}}\,,$$

provided that $\sup_{\xi'} |M(\xi',\boldsymbol{\theta})| > 0$. We assume that the set of $\boldsymbol{\theta} \in \mathbb{R}^p$ such that $\sup_{\xi'} |M(\xi',\boldsymbol{\theta})| = 0$ is of measure zero with respect to the prior distribution. This set is the the collection of singularities if M is continuous. This regularity condition holds in all of the examples we have considered, and it is sufficient to guarantee that the efficiency distribution is well-defined. It does not matter what value we assign to $\mathrm{eff}(\xi|\boldsymbol{\theta})$ for singular $\boldsymbol{\theta}$ since this only affects events of probability zero.

From this perspective, the optimisation of an objective function which involves an average over our prior beliefs on $\boldsymbol{\theta}$ is a device to aid us in obtaining a satisfactory efficiency distribution by producing a 1-dimensional summary of this distribution. This is essentially a pseudo-Bayesian viewpoint which assumes that the resulting analysis will be performed in a non-Bayesian fashion.

Chaloner and Verdinelli (1995) discourage the use of efficiency-based criteria, citing among other things the lack of a canonical choice of such a criterion. However, in those cases where the principles behind the justification of (7.2) break down, and (7.3) is degenerate, it seems sensible to make use of a criterion which is well-behaved. The choice between different efficiency-based criteria can perhaps be made by the extent to which they penalise very low efficiencies in certain parameter scenarios.

7.4.2 Mean local efficiency

An alternative to the optimisation of (7.3) is to maximise the *mean local efficiency*,

$$\Psi(\xi) = E_{\boldsymbol{\theta}}\, \mathrm{eff}(\xi|\boldsymbol{\theta})\,, \tag{7.14}$$

which is well-defined and satisfies

$$0 \le \Psi(\xi) \le 1\,,$$

for all choices of model, prior distribution, and design such that the set of $\boldsymbol{\theta} \in \mathbb{R}^p$ with $\sup_{\xi'} |M(\xi',\boldsymbol{\theta})| = 0$ is of measure zero. A pseudo-decision-theoretic justification of this criterion, assuming that the analyst will use non-Bayesian methods, is given in Section 7.5.2, together with practical issues surrounding the computation of (7.14). The objective function can be shown to be concave on the space of design measures, and also differentiable. Therefore a General Equivalence Theorem holds for this criterion.

7.4.3 Firth and Hinde's I_α

Another possible alternative criterion is to maximise the objective function of Firth and Hinde (1997),

$$I_\alpha(\xi) = \begin{cases} (1/\alpha)\log[E_{\boldsymbol{\theta}}\{|M(\xi,\boldsymbol{\theta})|^\alpha\}] & \alpha \neq 0 \\ E_{\boldsymbol{\theta}}\log\{|M(\xi,\boldsymbol{\theta})|\} & \alpha = 0\,. \end{cases} \tag{7.15}$$

Those authors show that I_α is concave and differentiable provided $\alpha \leq 1/p$, where p is the dimension of M. In this case, the criterion therefore satisfies a General Equivalence Theorem. There are no issues with convergence for this criterion for $0 < \alpha \leq 1/p$, for details and further issues see Section 7.5.1.

7.4.4 Modification of prior

If one adjusts the prior distribution, bounding its support away from any singularities, then the objective function (7.3) will converge. The resulting optimal design can be assessed in terms of the efficiency distribution using the 'true', unadjusted prior distribution.

Note that making this modification causes the criterion to totally ignore the performance of the design in the neighbourhood of the singularities. Thus if the practitioner is seriously concerned that the parameters may be arbitrarily close to singular values, another approach should be used.

7.4.5 Density designs

With (7.1) we defined a design to be a (finitely-supported) discrete probability measure on $\mathcal{X}$. Sometimes it is beneficial to generalise this by allowing a design to be an arbitrary probability measure on $\mathcal{X}$. In particular, this allows us to define a design in terms of a probability density function, $g(\mathbf{x})$, on $\mathcal{X}$. We refer to these as *density designs.* In Section 7.6.2 we give an example where all finitely supported designs are singular with respect to (7.3), but there are nonsingular density designs.

For an arbitrary design measure ξ, the information matrix satisfies

$$M(\xi,\boldsymbol{\theta}) = \int_{\mathcal{X}} M(\mathbf{x},\boldsymbol{\theta})\, d\xi(\mathbf{x})\,.$$

When ξ is finitely-supported, this is

$$M(\xi,\boldsymbol{\theta}) = \sum_{i=1}^{k} w_i M(\mathbf{x}_i,\boldsymbol{\theta})\,,$$

and for density designs it is

$$M(\xi,\boldsymbol{\theta}) = \int_{\mathcal{X}} M(\mathbf{x},\boldsymbol{\theta}) g(\mathbf{x}) d\mathbf{x}\,.$$

The latter can be used to evaluate the objective function (7.3), via

$$E_{\boldsymbol{\theta}} \log |M(\xi, \boldsymbol{\theta})| = \int_{\mathbb{R}^p} \log \left| \int_{\mathcal{X}} M(\mathbf{x}, \boldsymbol{\theta}) g(\mathbf{x}) d\mathbf{x} \right| f(\boldsymbol{\theta}) d\boldsymbol{\theta} \,.$$

A density design is not implementable without further consideration. To create a design we can use in practice, we draw a finite sample from this measure. The properties of this scheme are studied in Section 7.7, but a key property is that the randomness of the sampled designs means there can always be a positive probability of obtaining a reasonably efficient design, even if on average the performance is worse than using a deterministic design. Note however that results on density designs must be obtained analytically, and in general this is intractable.

There is clearly a parallel to drawn between the situation here and that discussed by Wiens (1992). The latter considers designs for linear regression which are robust to functional departures from the assumed model within various classes. When the widest class of alternatives is entertained, specifically an L^2-neighbourhood of 0, it is found that all finitely supported designs are singular with respect to the minimax criterion, while designs defined by a density function are non-singular.

7.5 Alternative optimality criteria

7.5.1 Firth and Hinde's I_α

In Section 7.4.3 we stated that there are no convergence issues with I_α for $0 < \alpha \leq 1/p$. Let us now clarify why this is the case. The key observation is the following.

Lemma 7.1. *The expectation involved in (7.15), $E_{\boldsymbol{\theta}}\{|M(\xi, \boldsymbol{\theta})|^\alpha\}$, is non-negative. Moreover it is equal to 0 if and only if $|M(\xi, \boldsymbol{\theta})| = 0$ for all $\boldsymbol{\theta}$. Thus ξ is singular with respect to I_α if and only if ξ is singular locally for all parameter values. In other words, $I_\alpha(\xi) > -\infty$ unless ξ is* uniformly *singular.*

The result essentially follows by continuity of the integrand. Thus we do not need be concerned with negative infinities. Provided we are willing to assume a mild regularity condition, we do not need to worry about positive infinities either.

Lemma 7.2. *A sufficient condition for $I_\alpha(\xi) < \infty$ for all ξ is that for any fixed ξ there do not exist parameter values such that $|M(\xi, \boldsymbol{\theta})|$ is arbitrarily large.*

A potential criticism of I_α is that it does not account for the scale (in other words, the maximum possible value) of $|M(\xi, \boldsymbol{\theta})|$ at a particular value of $\boldsymbol{\theta}$. Thus it may be possible to make a large impact on the value of the objective function by focussing efforts on the performance of ξ at values of $\boldsymbol{\theta}$ where $\sup_{\xi'} |M(\xi', \boldsymbol{\theta})|$ is large. We would ideally like a scale-free objective function which rewards designs that perform 'as well as possible' for different possible true values of $\boldsymbol{\theta}$, with weighting corresponding to the probability of those values. This is indeed provided by the mean local efficiency, (7.14).

7.5.2 Mean local efficiency

The mean local efficiency criterion can be justified in a pseudo-decision theoretic way as follows. We assume equal financial cost per run, and that the (frequentist) objective of the experiment is to produce a confidence interval for the parameters.

Let c be the budget for the experiment. Suppose we spend the entire budget running an experiment with a design which is $100e\%$ efficient. If we had made a better choice and used the optimal design instead, we would have spent a smaller amount, namely ec, while expecting to obtain confidence intervals of the same size. Thus the financial value of the information we obtain from our inefficient experiment is ec, which is the cost to run the cheapest equally informative experiment. We can thus regard the loss due to inefficiency as

$$L(\xi, \boldsymbol{\theta}) = c(1 - \mathrm{eff}(\xi|\boldsymbol{\theta}))\,.$$

Choosing ξ to minimise our expected loss with respect to the prior distribution on $\boldsymbol{\theta}$ is equivalent to maximising the mean local efficiency.

Note that the above justification assumes that the prior distribution will not be used in the analysis. This might be the case if: (i) it is difficult to elicit a prior which accurately summarises expert beliefs, thus we might use a 'rough' prior to design the experiment, (ii) the designer and the analyst have differing prior beliefs about $\boldsymbol{\theta}$, or (iii) an objective analysis not incorporating prior beliefs is required.

As we stated in Section 7.4.2, the objective function (7.14) is concave. Moreover it is differentiable, and the derivative of Ψ at ξ_2 in the direction of ξ_1 is

$$\psi(\xi_2, \xi_1) = \frac{1}{p} E_{\boldsymbol{\theta}}\{\mathrm{eff}(\xi_2|\boldsymbol{\theta})\,\mathrm{tr}[M(\xi_1, \boldsymbol{\theta}) M(\xi_2, \boldsymbol{\theta})^{-1}]\} - E_{\boldsymbol{\theta}}\{\mathrm{eff}(\xi_2|\boldsymbol{\theta})\}\,. \tag{7.16}$$

For proofs, see Section 7.9.2.

In practice, the expectation (7.14) must be evaluated numerically. This may be via Monte Carlo or a quadrature scheme, such as that of Gotwalt et al. (2009). In the case of numerical quadrature, we estimate (7.14) by

$$\Psi(\xi) \approx \sum_{i=1}^{n_a} \gamma_i\, \mathrm{eff}(\xi|\boldsymbol{\theta}_i)\,,$$

where $\boldsymbol{\theta}_i$ and γ_i, $i = 1, \ldots, n_a$, are the integration abscissae and weights respectively. To evaluate $\mathrm{eff}(\xi|\boldsymbol{\theta}_i)$ we must in general find the locally optimal design at $\boldsymbol{\theta}_i$ using numerical maximisation. The fact that we cannot be completely certain to have found the optimal design means that, in general, our numerical approximations to the local efficiencies will be overestimates. It is computationally more straightforward to approximate the mean local efficiency if there is an analytical form for the locally optimal designs, such as in Chaloner and Larntz (1989) or Russell, Woods, Lewis and Eccleston (2009).

7.6 Analytical results for exponential model

We now turn our attention to the exponential model (7.7). Proofs of the results given in Section 7.6–7.6.2 are presented in Section 7.9.5. The information matrix for a finitely supported design ξ is

$$\begin{aligned} M(\xi, \theta) &= \sum_{i=1}^{k} w_i \left(\frac{\partial \eta(x_i, \theta)}{\partial \theta} \right)^2 \\ &= \sum_{i=1}^{k} w_i \left(\frac{x_i^2}{\theta^4} \right) e^{-2x_i/\theta} \,. \end{aligned}$$

From this it can be seen that there are singularities at $\theta = 0, \infty$. Note however that for $\theta \in (0, \infty)$ the only singular design is that which places unit mass at $x = 0$. All other single point designs are locally nonsingular for $\theta > 0$.

Let us assume a $U(0, a)$ prior distribution, $a > 0$, which presupposes that θ may be arbitrarily close to the singularity at 0. Then we have the two rather dramatic results:

Lemma 7.3. *All single-point designs are ϕ-singular.*

Theorem 7.1. *All finitely supported designs are ϕ-singular.*

The proof of the first result proceeds by direct integration, and for the second we approximate $\phi(\xi)$ by $\phi(\min_i x_i)$ as the design point closest to 0 can be shown to dominate the integral. For full details see Section 7.9.5. Note that trivially these results hold also for the alternative approximation (7.4) since the prior is flat, $R = d^2 \log f / d\theta^2 = 0$.

These results essentially state that the range of scenarios entailed in the $U(0, a)$ prior is too broad for there to be a satisfactory trade-off under ϕ using finitely supported designs. Nonetheless, finitely-supported optimal designs can be obtained under alternative criteria such as the mean local efficiency Ψ, as in Section 7.6.1. If we are determined to continue to look at ϕ, we must consider designs defined by probability density functions. We will show in Section 7.6.2, there exists such a density design which is nonsingular with respect to ϕ.

In fact, conditions on any smooth prior density supported on $[0, a]$ can be obtained for ϕ to be non-degenerate on the set of finitely supported designs.

Lemma 7.4. *A necessary condition for there to exist ϕ-nonsingular finitely supported designs is that the prior density function $f(\theta)$ satisfies $f(0) = 0$. This condition is also sufficient.*

The proof proceeds by Taylor series expansion of the prior density function at $\theta = 0$.

7.6.1 Optimal designs

In the exponential model, it is straightforward to obtain locally optimal designs using simple calculus.

Lemma 7.5. *The locally optimal design at θ is the single point design $x = \theta$. The maximal value of $M(\xi, \theta)$ is*

$$\sup_{x>0} M(x, \theta) = \frac{1}{e^2 \theta^2} \,.$$

Note as a corollary that

$$\text{eff}(x|\theta) = \frac{e^2x^2}{\theta^2}e^{-2x/\theta}\,.$$

Furthermore we can obtain the Ψ-optimal design analytically.

Lemma 7.6. *The design which maximises the mean local efficiency (7.14) under $\theta \sim U(0,a)$ is that which assigns unit mass to $x = a/2$. Moreover the mean efficiency in this case is 67%, irrespective of the value of a.*

The local efficiency of the Ψ-optimal design is plotted as a function of θ in Figure 7.2, in the case $a = 5$. Note the extremely poor performance in the region of $\theta = 0$, and local optimality for $\theta = a/2$. Figure 7.3 shows the probability density function for the local efficiency distribution induced by the prior distribution on θ. This plot was obtained by considering $\text{eff}(x|\theta)$ as a transformation of θ, and numerically inverting the efficiency function (for more details, see Section 7.9.6). Note in particular the infinite density at $\text{eff}(\xi,\theta) = 0$ and the jump around $\text{eff}(\xi,\theta) = 0.7$. The high density of efficiencies near 0 is due to the flatness of the efficiency function in this region, and the jump occurs near 0.7 because this is the break point above which there are two values of θ which give rise to the same efficiency.

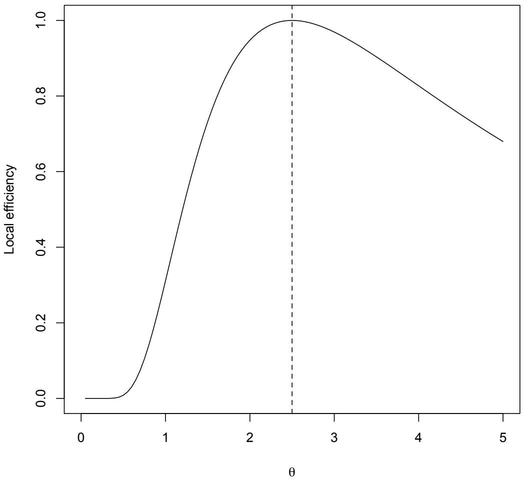

Figure 7.2: Exponential model: local efficiency of Ψ-optimal design, $0 \leq \theta \leq a$

7.6.2 Density designs

First note that the exponential model has only one parameter, therefore $M(x,\theta)$ is a scalar and we can show that

$$\text{eff}(\xi|\theta) = \int_{\mathcal{X}} \text{eff}(x|\theta)g(x)dx\,.$$

We obtain the following results for the uniform design defined by the probability density function $g(x) = a^{-1}$, $0 < x < a$.

Lemma 7.7. *The uniform design, $\xi \sim U(0,a)$, is ϕ-nonsingular.*

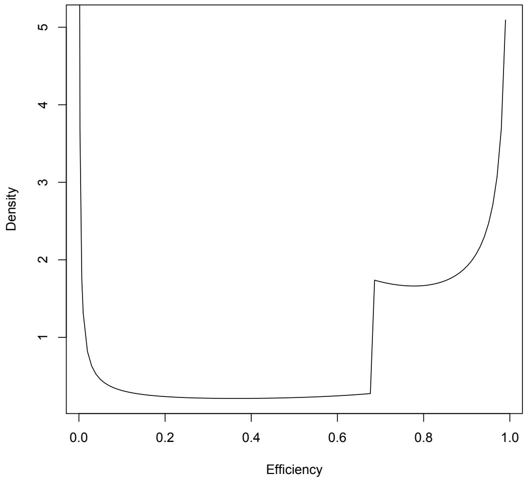

Figure 7.3: Exponential model: local efficiency density, Ψ-optimal design

Lemma 7.8. *The efficiency* $\mathrm{eff}(\xi|\theta)$ *of the uniform design is analytically tractable. Moreover, as* $\theta \to 0$, *it is*

$$\mathrm{eff}(\xi|\theta) \sim \frac{e^2}{4a}\theta\,.$$

For the case $a = 5$, the local efficiency of the uniform design is plotted as a function of θ in Figure 7.4, together with the limiting behaviour given in the previous Lemma. Note that the efficiency for small θ is much better than with the Ψ-optimal design, although the practical significance is only made clear in Section 7.7. For small efficiencies, $\mathrm{eff}(\xi, \theta)$ is essentially a linear transformation of θ and so the distribution of small efficiencies will be close to uniform. As a result, there is no peak at 0 in the local efficiency density. For a plot of this efficiency density, see Figure 7.5.

We make no claim that the uniform design is optimal. Derivation of an optimal density in this case would be a formidable task as a result of the form of the objective function. However, we can obtain a lower bound on the 'Bayesian' efficiency of the design, which we define as

$$\text{Bayes-eff}(\xi) = 100 \times \exp\{\phi(\xi) - \sup_{\xi'} \phi(\xi')\}\,\%\,. \tag{7.17}$$

The definition of the quantity (7.17) is justified as follows. First note that

$$\phi(\xi) = E_\theta \log M(\xi, \theta) = E_\theta \log \mathrm{eff}(\xi|\theta) + E_\theta \log s(\theta)\,,$$

where $s(\theta) = \sup_{\xi'} M(\xi', \theta)$ is the maximal value of the local objective function for a particular θ. As $s(\theta)$ does not depend on the chosen design, (7.17) can be rewritten as

$$\text{Bayes-eff}(\xi) = 100 \times \exp\{\phi_E(\xi) - \sup_{\xi'} \phi_E(\xi')\}\%\,,$$

with $\phi_E(\xi) = E_\theta \log \mathrm{eff}(\xi|\theta)$.

Let ξ_1, ξ_2 be density designs and $\phi_E(\xi_i) = \log e_i$, $i = 1, 2$. Then ξ_1 is equivalent, modulo ϕ_E to a (hypothetical) design, ξ_{e_1}, which is $100e_1\%$ efficient for all θ. Similarly, modulo ϕ_E we have that ξ_2 is equivalent to a design, ξ_{e_2}, which is $100e_2\%$ efficient for all θ. Clearly the only possibility for $\text{eff}(\xi_{e_1}|\xi_{e_2})$ is $e_1/e_2 \times 100\%$. By the equivalences we have stated, so too $\text{eff}(\xi_1|\xi_2) = e_1/e_2 \times 100\%$.

Lemma 7.9. *The uniform design has* Bayes-eff$(\xi) \geq 69\%$ *independently of* a.

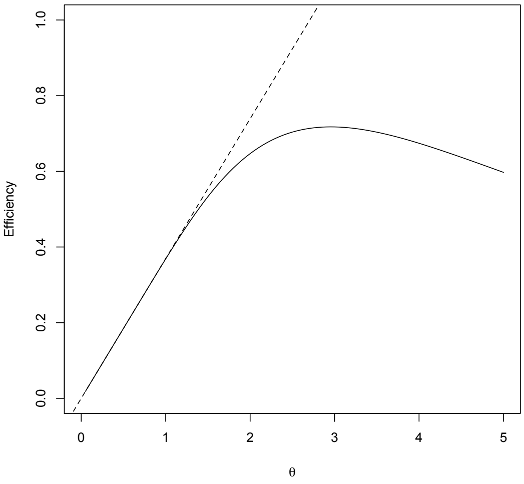

Figure 7.4: Exponential model: local efficiency function for the uniform design, $\xi \sim U(0, a)$, $0 \leq \theta \leq a$

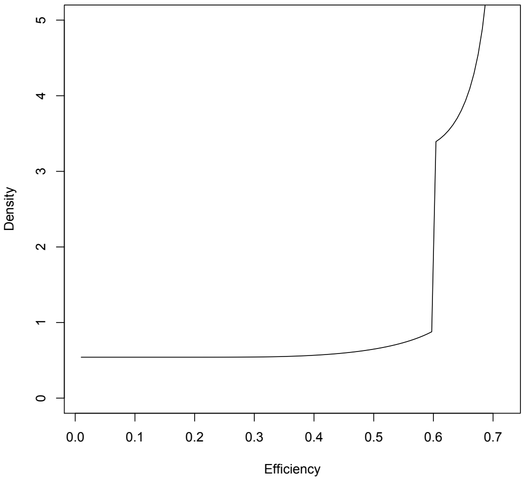

Figure 7.5: Exponential model: probability density of local efficiency distribution, uniform design

7.7 Interpretation of density designs

In this section we study the properties of finite samples drawn from the design density $\xi = U(0,a)$ for the exponential model (7.7) with prior $\theta \sim U(0,a)$. Let $X_n = (x_1, \ldots, x_n)^T$ be such a sample. Then $E_\theta \log |M(X_n;\theta)| = -\infty$ with certitude, therefore there is no sense in which the value of the objective function at X_n converges to that at ξ. However, the efficiency of X_n satisfies certain desirable asymptotic properties as $n \to \infty$.

7.7.1 Asymptotic properties

Considering efficiencies we observe that

$$\mathrm{eff}(X_n|\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{eff}(x_i|\theta) \tag{7.18}$$

$$E_{X_n}\{\mathrm{eff}(X_n|\theta)\} = \mathrm{eff}(\xi|\theta)\,.$$

Theorem 7.2. *By the strong law of large numbers, for all θ,*

$$\mathrm{eff}(X_n|\theta) \to \mathrm{eff}(\xi|\theta) \quad \textit{almost surely as } n \to \infty\,. \tag{7.19}$$

In other words the efficiency of the sampled design converges to that of the density design.

Note that (7.18) is only true because we are in a 1-parameter model, however the conclusion (7.19) can most likely be obtained when there are more parameters by considering convergence of the information matrix.

Moreover, note that as the efficiency is a sum of IID random variables, we can apply the central limit theorem to show that, for large n, $\mathrm{eff}(X_n|\theta)$ is approximately normally distributed with mean $\mathrm{eff}(\xi|\theta)$ and variance

$$v(\theta) = \frac{1}{n}\left(E_x[\mathrm{eff}(x|\theta)^2] - \mathrm{eff}(\xi|\theta)^2\right)\,,$$

which can be computed analytically. This can be used to obtain approximate 95% performance limits, such as in Figure 7.6.

It can be shown using Lemma 7.8 that, as $\theta \to 0$, we have

$$E_x[\mathrm{eff}(x|\theta)^2] \sim \frac{3e^4}{128a}\theta\,,$$

so that, for small θ, the variance of the efficiency is approximately

$$\mathrm{var}(\mathrm{eff}(X_n|\theta)) \approx \frac{1}{n}\left(\frac{3e^4}{128a}\theta - \frac{e^4}{16a^2}\theta^2\right)\,.$$

Thus the performance limits become

$$\frac{e^2}{4a}\theta \pm 1.96\sqrt{\frac{1}{n}\left(\frac{3e^4}{128a}\theta - \frac{e^4}{16a^2}\theta^2\right)}\,.$$

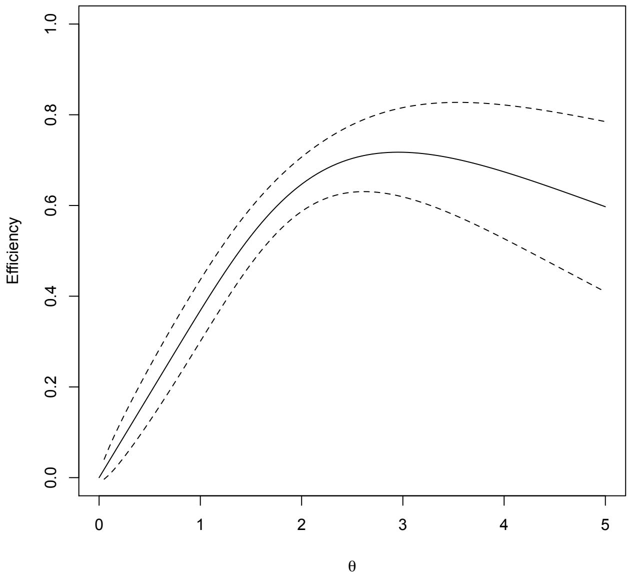

Figure 7.6: Approximate 95% performance limits for finite design of size $n = 100$ from the uniform design

Since the centre of the interval is $O(\theta)$ and the ± term is $O(\sqrt{\theta})$, if n is fixed then for small θ the lower limit is negative and hence not useful.

This seems to be saying that to guarantee good performance for all θ, we need to sample infinitely many points from ξ, which agrees with our intuition.

7.7.2 Benefits

The benefit of using a random sample from a design density is that no matter what the value of θ, there is always a positive chance that we will obtain a reasonably efficient design. This must of course be traded off against the positive probability of obtaining a highly inefficient design.

Let us fix θ, and a desired level of efficiency, say 70%. Note that

$$\text{eff}(x|\theta) = h(x/\theta)\,,$$

where

$$h(u) = e^2u^2e^{-2u}\,,$$

and so $\text{eff}(x|\theta) > 0.7$ if and only if $\gamma_1 \leq x/\theta \leq \gamma_2$, where $\gamma_1 < \gamma_2$ are the two roots of $h(\gamma) = 0.7$. Using this we can obtain a crude lower bound for the probability that $\text{eff}(X_n|\theta) \geq 0.7$ as follows

$$\begin{aligned} P(\text{eff}(X_n|\theta) \geq 0.7) &= P\left(\frac{1}{n}\sum_{i=1}^{n}\text{eff}(x_i|\theta) \geq 0.7\right) \\ &\geq P\left(\text{eff}(x_i|\theta) \geq 0.7 \text{ for } i = 1, \ldots, n\right) \end{aligned}$$

$$
\begin{aligned}
&\geq \prod_{i}^{n} P(\gamma_1 \theta \leq x_i \leq \gamma_2 \theta) \\
&\geq \theta^n (\gamma_2 - \gamma_1)^n \, .
\end{aligned}
$$

7.8 Discussion

In this chapter we have highlighted the potential impact of singularities on the convergence of the mean-log-determinant objective function, ϕ. In particular we have shown that if the range of prior parameter uncertainty is sufficiently wide that we approach singularities, it may be necessary either to abandon ϕ or to give up on finitely supported designs.

It can be a relatively difficult analytical problem even to identify singularities, let alone determine the rate of prior decay in their neighbourhood necessary to ensure that $\phi > -\infty$. Thus, in these cases of extreme parameter uncertainty it may be more practical to use an alternative criterion such as the mean local efficiency criterion.

It will be difficult in general problems to compute $\phi(\xi)$ for a density design, since only analytical methods will be able to determine if the objective function converges. However, in cases where it is possible to find nonsingular density designs analytically, it would be interesting to see whether ϕ-optimal density designs can be derived. It is not immediately obvious how one should extend the General Equivalence Theorem or the Federov-Wynn algorithm to density designs. To see this, consider taking the directional derivative of ϕ at ξ in the direction of a single design point x: by taking a convex combination of ξ with the single point design x one obtains a design which is not a density design and so not relevant to the question of optimality in the class of density designs.

Other future work could focus more directly on designing the local efficiency distribution, for instance penalising distributions with a comparatively large variance or having a high probability attached to very low efficiencies.

7.9 Appendix: Proofs and further analytical results

7.9.1 Singularity property of Firth and Hinde's I_α

Proof of Lemma 7.1. We assume the regularity condition that $M(\xi, \boldsymbol{\theta})$ is a continuous function of ξ and $\boldsymbol{\theta}$. Observe that $I_\alpha = -\infty$ precisely when $\int_{\boldsymbol{\theta}} |M(\xi;\boldsymbol{\theta})|^\alpha f(\boldsymbol{\theta}) d\boldsymbol{\theta} = 0$, where f is the (continuous) prior density function. As M is non-negative definite, $|M|^\alpha \geq 0$. Thus $|M|^\alpha f(\boldsymbol{\theta})$ is non-negative and continuous and from results in mathematical analysis, $\int_{\boldsymbol{\theta}} |M(\xi;\boldsymbol{\theta})|^\alpha f(\boldsymbol{\theta}) d\boldsymbol{\theta} = 0$ implies $|M|^\alpha f(\boldsymbol{\theta}) = 0$ for all $\boldsymbol{\theta} \in \Theta$. The latter condition can occur only if $|M(\xi;\boldsymbol{\theta})| = 0$ for all $\boldsymbol{\theta}$ in the support of f. This proves the 'only if' implication, the 'if' part is trivial. □

Lemma 7.10. *Suppose that $M(\xi;\boldsymbol{\theta})$ is continuous function of ξ and $\boldsymbol{\theta}$. Then $\Psi(\xi) = E_{\boldsymbol{\theta}}(\mathrm{eff}(\xi|\boldsymbol{\theta})) = 0$ iff $|M(\xi;\boldsymbol{\theta})| = 0$ for all $\boldsymbol{\theta}$.*

The proof is analogous to the above.

7.9.2 Concavity and differentiability of mean local efficiency

We use equation (4) of Firth and Hinde (1997),

$$|\delta M_1 + (1-\delta)M_2|^\alpha \geq \delta |M_1|^\alpha + (1-\delta)|M_2|^\alpha \, . \tag{7.20}$$

which holds for $\alpha \leq 1/p$, where p is the dimension of M. To demonstrate that $\Psi(\xi) = E_{\boldsymbol{\theta}}(\mathrm{eff}(\xi|\boldsymbol{\theta}))$ is concave, set $\alpha = 1/p$ and let $M_1 = M(\xi_1, \boldsymbol{\theta})$ and $M_2 = M(\xi_2, \boldsymbol{\theta})$ in (7.20). Then divide the inequality through by $|M^*|^{1/p} = \sup_{\xi'} |M(\xi', \boldsymbol{\theta})|^{1/p}$ and take expectations with respect to $\boldsymbol{\theta}$. This yields

$$E_{\boldsymbol{\theta}}\left(\frac{|M(\delta\xi_1 + (1-\delta)\xi_2, \boldsymbol{\theta})|^{1/p}}{|M^*|^{1/p}}\right) \geq \delta E\left(\frac{|M_1|^{1/p}}{|M^*|^{1/p}}\right) + (1-\delta)E\left(\frac{|M_2|^{1/p}}{|M^*|^{1/p}}\right) . \tag{7.21}$$

Recall the definition of mean efficiency,

$$\Psi(\xi) = E(\mathrm{eff}(\xi|\boldsymbol{\theta})) = E\left\{\frac{|M(\xi, \boldsymbol{\theta})|^{1/p}}{\sup_{\xi'} |M(\xi', \boldsymbol{\theta})|^{1/p}}\right\} .$$

Hence, using (7.21),

$$E\,\mathrm{eff}(\delta\xi_1 + (1-\delta)\xi_2|\boldsymbol{\theta}) \geq \delta E\,\mathrm{eff}(\xi_1|\boldsymbol{\theta}) + (1-\delta)E\,\mathrm{eff}(\xi_2|\boldsymbol{\theta}) \, ,$$

and so Ψ is concave as claimed.

The derivative of Ψ at ξ_2 in the direction of ξ_1 is defined as

$$\begin{aligned}\psi(\xi_2, \xi_1) &= \lim_{\delta \to 0} \delta^{-1}\{\Psi\big((1-\delta)\xi_2 + \delta\xi_1\big) - \Psi(\xi_2)\} \\ &= \frac{d}{d\delta}\bigg|_{\delta=0} \Psi\big((1-\delta)\xi_2 + \delta\xi_1\big) \, .\end{aligned}$$

To calculate this derivative, let us first define the shorthand

$$M = \delta M_1 + (1-\delta)M_2 \, ,$$

and note that (Silvey, 1980, p.21)

$$\frac{d}{d\delta}\bigg|_{\delta=0} \log|M| = \mathrm{tr}(M_1 M_2^{-1}) - p \, .$$

Using the chain rule we calculate the derivative of $|M|^{1/p}$ as,

$$\begin{aligned}\frac{d}{d\delta}\bigg|_{\delta=0} |M|^{1/p} &= \frac{d}{d\delta}\bigg|_{\delta=0} \exp\{p^{-1}\log|M|\} \\ &= p^{-1}|M_2|^{1/p}\{\mathrm{tr}(M_1 M_2^{-1}) - p\} \, .\end{aligned}$$

As $|M^*|$ does not depend on δ, we can obtain the derivative of the local efficiency by dividing the above through by $|M^*|^{1/p}$,

$$\frac{d}{d\delta}\bigg|_{\delta=0} \left\{\frac{|M|^{1/p}}{|M^*|^{1/p}}\right\} = p^{-1}\,\mathrm{eff}(\xi_2|\boldsymbol{\theta})\{\mathrm{tr}(M_1 M_2^{-1}) - p\} \, .$$

Finally, it is straightforward to obtain the directional derivative of the mean local efficiency, we differentiate under the expectation sign to obtain

$$\begin{aligned}\psi(\xi_2, \xi_1) &= \frac{d}{d\delta}\Big|_{\delta=0} E\left\{\frac{|M|^{1/p}}{|M^*|^{1/p}}\right\} \\ &= E\left\{\frac{d}{d\delta}\Big|_{\delta=0} \frac{|M|^{1/p}}{|M^*|^{1/p}}\right\} \\ &= \frac{1}{p}E\{\text{eff}(\xi_2|\boldsymbol{\theta})\,\text{tr}(M_1 M_2^{-1})\} - E\{\text{eff}(\xi_2|\boldsymbol{\theta})\}\,.\end{aligned}$$

7.9.3 Singularity in logistic model at $\beta = \infty$

We must consider the case $x = \mu$ slightly differently and so we change the notation for a design, writing instead

$$\Xi = \left\{\begin{array}{cccc} \mu & x_1 & \cdots & x_k \\ w_\mu & w_1 & \cdots & w_k \end{array}\right\},$$

where for all i, $x_i \neq \mu$. As μ is the dose at which $p = 1/2$, $\lambda_\mu = 1/4$ and we have that

$$|M(\Xi, \boldsymbol{\theta})| = \beta^2 \left((1/4)w_\mu + \sum_{i=1}^{k} w_i\lambda_i\right)\left((1/4)w_\mu(\mu - \bar{x})^2 + \sum_{i=1}^{k} w_i\lambda_i(x_i - \bar{x})^2\right). \tag{7.22}$$

Note that

$$\begin{aligned}\beta^2\lambda_i &= \frac{\beta^2 e^{\beta|x_i-\mu|}}{(1 + e^{\beta|x_i-\mu|})^2} \\ &\to 0 \text{ as } \beta \to \infty\,,\end{aligned}$$

which is sufficient to establish that

$$\lim_{\beta\to\infty} |M(\Xi, \boldsymbol{\theta})| = \begin{cases} 0 & \text{if } w_\mu = 0 \\ (1/16)w_\mu^2 \lim_{\beta\to\infty} \beta^2(\mu - \bar{x})^2 & \text{if } w_\mu > 0\,, \end{cases} \tag{7.23}$$

as when we expand (7.22) all other terms contain factors of the form $\beta^2\lambda_i$ and hence vanish in the limit. We now consider the case $w_\mu > 0$. Note that,

$$\begin{aligned}\beta(\mu - \bar{x}) &= \frac{\sum_{i=1}^{k} w_i(\lambda_i\beta)(\mu - x_i)}{(1/4)w_\mu + \sum_{i=1}^{k} w_i\lambda_i} \\ &\to 0\,,\end{aligned}$$

as $\beta \to \infty$ (and therefore also $\bar{x} \to \mu$). This follows since $\beta\lambda_i \to 0$ and $\lambda_i \to 0$, therefore the numerator converges to 0 and the denominator converges to $(1/4)w_\mu > 0$. Therefore from (7.23) it is clear there is a singularity at $\beta = \infty$.

7.9.4 Binary GLMM

Here we demonstrate formally the limiting property of the model which concerns us. As the likelihood and information matrix are defined in terms of integrals with respect to u, we require results from measure theory which allow us to compute the limits of sequences of integrals. Notably we use Lebesgue's dominated convergence theorem, which is stated fully in Theorem 3.2. We apply the result below to the case that $\sigma^2 \to \infty$.

First, we recall some notation from Chapters 2 and 3. Let m be the block size, and $\zeta = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \subset [-1,1]^q$ be an arbitrary block. Also let $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_m)^T$ the the corresponding vector of the fixed parts of the linear predictors, $\eta_j = \mathbf{f}^T(\mathbf{x}_j)\boldsymbol{\beta}$, $j = 1, \ldots, m$. Then $p_{\mathbf{y}}(\boldsymbol{\eta}, \sigma^2)$ is the probability of obtaining responses $\mathbf{y}$ from block ζ. Recall also that h denotes the logistic function, $h(\eta) = 1/(1+e^{-\eta})$, and ϕ_{σ^2} is the density function of a $N(0, \sigma^2)$ random variable.

Lemma 7.11. *As $\sigma^2 \to \infty$, the logistic random intercept model is degenerate in the way described in Section 2.1. Namely,*

$$p_{\mathbf{y}}(\boldsymbol{\eta}, \sigma^2) \to \begin{cases} 1/2 & \textit{for } \mathbf{y} = \mathbf{1} = (1,1,\ldots,1)^T \textit{ and } \mathbf{y} = \mathbf{0} = (0,0,\ldots,0)^T \\ 0 & \textit{for all other } \mathbf{y} \in \{0,1\}^m. \end{cases}$$

Proof. Let $\zeta = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ be an arbitrary block. For $\sigma^2 > 0$, define

$$G_{\sigma^2}(u) = \prod_{i=1}^{m} h(\eta_j + \sigma u)\phi_1(u)\,. \tag{7.24}$$

Recall from (3.10) that $p_{\mathbf{1}}(\boldsymbol{\eta}, \sigma^2) = \int_{\mathbb{R}} G_{\sigma^2}(u)du$. As our measure space we take $(\mathbb{R}, \mathcal{B}, \mathcal{L})$, where $\mathcal{B}$ is the Borel σ-algebra on $\mathbb{R}$, and $\mathcal{L}$ is Lebesgue measure. Since $0 \le h(\eta) \le 1$ for all $\eta \in \mathbb{R}$, the function G_{σ^2} is dominated by the integrable function ϕ_1. Moreover as $\sigma^2 \to \infty$, G_{σ^2} converges pointwise to the function G_∞, given by

$$G_\infty(u) = \begin{cases} \phi_1(u) & \text{for } u > 0 \\ 0 & \text{for } u < 0\,. \end{cases}$$

The above limit can be verified by noting that, if $u > 0$ then $h(\eta_j + \sigma u) \to 1$ as $\sigma \to \infty$, whereas if $u < 0$ then $h(\eta_j + \sigma u) \to 0$.

Let $\sigma_n^2 > 0$ be an arbitrary positive sequence such that $\sigma_n^2 \to \infty$ as $n \to \infty$. By dominated convergence, as $n \to \infty$, $p_{\mathbf{1}}(\boldsymbol{\eta}, \sigma_n^2) = \int_{\mathbb{R}} G_{\sigma_n^2}(u)du \to \int_{\mathbb{R}} G_\infty(u)du = 1/2$. This is sufficient to establish that as $\sigma^2 \to \infty$, $p_{\mathbf{1}}(\boldsymbol{\eta}, \sigma^2) \to 1/2$, since if this were not the case we could construct a sequence of σ_n^2 with $\sigma^2 \to \infty$ and $p_{\mathbf{1}}(\boldsymbol{\eta}, \sigma_n^2)$ not converging to 1/2 (contradicting the above).

An analogous argument shows that as $\sigma^2 \to \infty$, $p_{\mathbf{0}}(\boldsymbol{\eta}, \sigma^2) \to 1/2$. Since probabilities sum to unity,

$$p_{\mathbf{0}} + p_{\mathbf{1}} + \sum_{\mathbf{y} \ne \mathbf{0}, \mathbf{1}} p_{\mathbf{y}} = 1\,.$$

Taking limits as $\sigma^2 \to \infty$,

$$(1/2) + (1/2) + \sum_{\mathbf{y} \ne \mathbf{0}, \mathbf{1}} \lim_{\sigma^2 \to \infty} p_{\mathbf{y}} = 1\,,$$

and so $\lim_{\sigma^2 \to \infty} p_{\mathbf{y}} = 0$ for $\mathbf{y}$ other than $\mathbf{1}$ or $\mathbf{0}$. □

7.9.5 Exponential model

Proof of Lemma 7.3 (ϕ-singularity of single point designs). Let $x \in (0, a)$. Then

$$\log |M(x, \theta)| = -\frac{2x}{\theta} - 4 \log \theta + 2 \log x \,.$$

Thus, if $\theta \sim U(0, a)$, $a > 0$, then

$$\begin{aligned} E_\theta(\log |M(x, \theta)|) &= \frac{1}{a} \int_0^a \left\{ -\frac{2x}{\theta} - 4 \log \theta + 2 \log x \right\} d\theta \\ &= \frac{1}{a} \left\{ \int_0^a -\frac{2x}{\theta} d\theta - \int_0^a 4 \log \theta d\theta + \int_0^a 2 \log x d\theta \right\} \\ &= -\infty - 4\{a \log a - a\}/a + 2a \log(x)/a \\ &= -\infty \,, \end{aligned}$$

due to the fact that $\int_0^a (1/\theta) d\theta = \infty$ and $\int_0^a \log \theta d\theta = a \log a - a$. Thus all one-point designs are ϕ-singular. □

Proof of Theorem 7.1 (ϕ-singularity of finitely supported designs). Let ξ be an arbitrary finitely supported approximate design, with support points and weights notated as in (7.9). Then

$$\log |M(\xi, \theta)| = \log \sum_{i=1}^{k} \frac{w_i x_i^2}{\theta^4} e^{-2x_i/\theta} \,.$$

Observe that we have the following: fix $x > 0$, then for all $y > 0$

$$\log(x + y) \leq \log(x) + y/x \,,$$

in other words the logarithm function is always below its tangents (easily established geometrically, or using calculus).

Thus we have that

$$\begin{aligned} \log |M(\xi, \theta)| &= \log \left\{ \sum_{i=1}^{k} \frac{w_i x_i^2}{\theta^4} e^{-2x_i/\theta} \right\} \\ &\leq \log \left\{ \frac{w_1 x_1^2}{\theta^4} e^{-2x_1/\theta} \right\} + \frac{\sum_{i=2}^{k} \frac{w_i x_i^2}{\theta^4} e^{-2x_i/\theta}}{\frac{w_1 x_1^2}{\theta^4} e^{-2x_1/\theta}} \\ &\leq \log \left\{ \frac{w_1 x_1^2}{\theta^4} e^{-2x_1/\theta} \right\} + \sum_{i=2}^{k} \frac{w_i x_i^2}{w_1 x_1^2} e^{-2(x_i - x_1)/\theta} \\ &\leq \log w_1 + \log M(x_1, \theta) + T(\theta) \,, \end{aligned} \tag{7.25}$$

where $T(\theta)$ is defined to be the term on the right. Without loss of generality we may reorder the x_i such that x_1 is the smallest, in other words $x_i - x_1 \geq 0$ for all i. This means that $0 \leq T(\theta) \leq \sum_{i=2}^{k} w_i x_i^2 / (w_1 x_1^2)$, and so $T(\theta)$ has a finite, positive mean with respect to the $U(0, a)$ distribution on θ.

However, as we have established in Lemma 7.3, the middle term of (7.25) has mean $-\infty$. Hence also $E(\log |M(\xi, \theta)|) = \log w_1 + E \log |M(x, \theta)| + E(T(\theta)) = -\infty$.

□

Proof of Lemma 7.4 (conditions on prior). Recall the result we wish to prove: Let the prior density function f be differentiable and supported on $[0, a]$. Then in the exponential model, a necessary and sufficient condition for $E(\log|M(x, \theta)|) > -\infty$ for all $x > 0$ is that $f(0) = 0$.

Combining this with Theorem 1, if $f(0) > 0$ then all finitely supported designs are singular with respect to the mean log-determinant.

(Sufficient) First note that

$$E_\theta(\log|M(x,\theta)|) = \frac{1}{a}\int_0^a \left\{-\frac{2x}{\theta} - 4\log\theta + 2\log x\right\} f(\theta)\, d\theta\,. \tag{7.26}$$

If f is differentiable at 0, with $f(0) = 0$, then by the definition of differentiability we can write

$$f(\theta) = \theta\{f'(0) + h(\theta)\}\,,$$

with h a function such that $h(\theta) \to 0$ as $\theta \to 0$.

In particular, given $K > 0$ there is $\epsilon > 0$ such that $h \leq K$ on $(0, \epsilon)$. Therefore, considering the first term in the integrand of (7.26), which is the only term which can cause us problems, over $(0, \epsilon)$ we have that

$$\begin{aligned}\int_0^\epsilon \frac{f(\theta)}{\theta} d\theta &\leq \int_0^\epsilon \frac{\theta\{f'(0) + K\}}{\theta} d\theta \\ &\leq \epsilon\{f'(0) + K\}\,.\end{aligned}$$

Clearly also

$$\int_\epsilon^a \frac{f(\theta)}{\theta}\, d\theta < \infty\,,$$

as the integrand is continuous, and therefore bounded, on the integration region. Hence the integral over the whole range $(0, a)$ is finite, ie

$$\int_0^a \frac{f(\theta)}{\theta} d\theta < \infty\,.$$

The second term in the integrand of (7.26) is unproblematic since f is clearly bounded on $(0, a)$. Therefore $\int_0^a f(\theta)\log\theta\, d\theta \leq \sup_\theta f(\theta) \int_0^a \log\theta\, d\theta < \infty$.

The third term in the integrand of (7.26) has finite integral provided $x > 0$.

(Necessary) Suppose $f(0) > 0$. Then f can be bounded below by some $L > 0$ on some interval $(0, \epsilon)$. Thus

$$\int_0^\epsilon \frac{f(\theta)}{\theta} d\theta \geq \int_0^\epsilon \frac{L}{\theta} d\theta = \infty\,.$$

This forces $E(\log|M(x, \theta)|) = -\infty$. □

Proof of Lemma 7.6 (optimal mean efficiency design). Note that

$$\begin{aligned}\text{eff}(x|\theta) &= \frac{x^2 e^2}{\theta^2} e^{-2x/\theta} \\ &= \frac{d}{d\theta}\left(\frac{xe^2}{2} e^{-2x/\theta}\right).\end{aligned}$$

Therefore when $\theta \sim U(0,a)$, we have that

$$E_\theta\{\text{eff}(x|\theta)\} = \frac{xe^2}{2a} e^{-2x/a}.$$

As we consider a 1-parameter model,

$$\begin{aligned}E_\theta\{\text{eff}(\xi|\theta)\} &= \sum_{i=1}^n w_i E_\theta\{\text{eff}(x_i|\theta)\} \\ &\le \sup_{x\in(0,a)} E_\theta\{\text{eff}(x|\theta)\},\end{aligned}$$

with equality when ξ is the design which assigns unit mass to $x_1 = \text{argmax}_{x\in(0,a)}\, xe^{-2x/a}$. It is easy to verify by calculus that $x_1 = a/2$. The mean efficiency of this design is

$$E_\theta\{\text{eff}\left(\frac{a}{2}\middle|\theta\right)\} = e/4 \approx 0.67.$$

□

Proof of Lemma 7.7 (ϕ-nonsingularity of $\xi = U(0,a)$, the uniform design). The information matrix is

$$\begin{aligned}M(\xi,\theta) &= \frac{1}{a}\int_0^a M(x,\theta)dx \\ &= \frac{1}{a\theta^4}\int_0^a x^2 e^{-2x/\theta}dx, \\ &= \frac{1}{a\theta^4}\left[-\frac{1}{4}\theta e^{-2x/\theta}(\theta^2 + 2\theta x + 2x^2)\right]_{x=0}^{x=a} \\ &= \frac{1}{4a\theta} - \frac{e^{-2a/\theta}(\theta^2 + 2\theta a + 2a^2)}{4a\theta^3}.\end{aligned} \tag{7.27}$$

Note that the first term on the RHS tends to ∞ as $\theta \to 0$, and the second term tends to 0. Thus in fact $M(\xi,\theta) \to \infty$ as $x \to 0$, and the uniform design can be made arbitrarily informative by taking θ sufficiently small.

To see that $M(\xi,\theta)$ has a positive minimum on $\theta \in (0,a)$, note that

$$\begin{aligned}M(\xi,\theta) &= \frac{1}{4a^2}H(\theta/a) \\ H(t) &= \{t^2 - e^{-2/t}[t^2 + 2t + 2]\}/t^3,\end{aligned}$$

where $H(t) : [0,1] \to \mathbb{R}$ can be checked to be monotone decreasing in t with value $H(1) \approx 0.323 > 0$. As $M(\xi,\theta) \ge 0.32/(4a^2)$, $\log|M|$ is bounded below and so $\phi(\xi) > -\infty$. □

Proof of Lemma 7.8 (efficiency of uniform design). To obtain an analytical expression for the efficiency, divide (7.27) by $\sup_{x\in(0,a)} M(x,\theta) = 1/(e^2\theta^2)$. To calculate the limiting behaviour as $\theta \to 0$, observe that the term containing $e^{-2a/\theta}$ must tend to 0. □

Proof of Lemma 7.9 (lower bound on Bayesian efficiency). Applying Jensen's inequality (to the concave function log) we obtain that

$$\begin{aligned}\phi_E(\xi) &= E_\theta \log \text{eff}(\xi|\theta)\\ &\le \log E_\theta E_x \, \text{eff}(x|\theta)\\ &\le \log E_x E_\theta \, \text{eff}(x|\theta)\,.\end{aligned}$$

By Lemma 7.8, the maximal mean efficiency for a single point design is $e/4 \approx 0.67$. Therefore, for all ξ,

$$\phi_E(\xi) \le 1 + \log(1/4)\,.$$

It is not immediately clear whether this upper bound is obtainable. However we use it to obtain a lower bound on the Bayesian efficiency of a given design. With ξ_u the uniform design, $\text{eff}(\xi_u|\theta) = \frac{e^2\theta^2}{4a^2} H(\theta/a) = G(\theta/a)$, and we have that

$$\begin{aligned}E_\theta \log \text{eff}(\xi_u|\theta) &= \int_0^a (1/a) \log G(\theta/a) d\theta\\ &= \int_0^1 \log G(t) dt\,,\end{aligned}$$

which is independent of a. We computed numerically $\int_0^1 \log G(t)dt \approx \log 0.465$. The Bayesian efficiency of the uniform design is therefore at least $0.465/0.67 \times 100\% \approx 69.5\%$, independently of a. □

7.9.6 Efficiency density plots

We can obtain better plots of the density function of the efficiency distribution by using the fact that $\text{eff}(\xi|\theta)$ is a transformation of the variable θ. Defining the shorthand

$$\mathcal{E}(\theta) = \text{eff}(\xi|\theta)\,,$$

and using square brackets notation for density functions, we have by a change of variable argument that

$$\begin{aligned}[\mathcal{E}](t) &= \sum_{\theta:\mathcal{E}(\theta)=t} \left|\frac{1}{\mathcal{E}'(\theta)}\right| f(\theta)\\ &= \sum_{\theta:\mathcal{E}(\theta)=t} \left|\frac{1}{a\mathcal{E}'(\theta)}\right|\,,\end{aligned}$$

where $f(\theta) = a^{-1}\mathbf{1}\{\theta \in (0,a)\}$ is the prior density function. This is analogous to the usual result for the density of a monotonic differentiable transformation of a random variable of known density. However here we must take into account that, for given $t \in [0,1]$, there may be multiple θ such that $\mathcal{E}(\theta) = t$.

Analytical expressions for $\mathcal{E}'$ are available in the examples under consideration, and we can find all the solutions to $\mathcal{E}(\theta) = t$ numerically. Thus we can obtain a direct numerical estimate of the efficiency density, without having to simulate from any distributions.

Chapter 8

Discussion and areas for development

8.1 Approximations for GLMMs

In Chapter 2 we developed the MQL and PQL approximations to the information matrix and applied them to calculate Bayesian designs for some GLMMs containing random intercepts. The designs resulting from these cheap analytical approximations were compared to designs from a more direct computational approximation, using MLNI, in Chapter 3. We have collected evidence which seems to suggest that in general MQL leads to more efficient designs than does PQL, for two reasons. First of all, PQL is not sufficiently sensitive to the allocation used when σ^2 is large (Section 3.1), and secondly it tends to find treatments which are worse than if we did not take into account the random effects at all (Sections 3.4 and 4.4.5).

One objective for future work is to calculate designs when the random effects structure is more complicated, for instance allowing the effects of the x_i to vary from block-to-block. MQL and PQL can be used in these situations at no extra cost, though our results suggest that PQL is unlikely to be a good choice. An obvious question is whether we are able to evaluate the performance of the approximations in this new situation. It would also be interesting in general to compare the performance of our approximations with the 'Monte Carlo PQL' of Tekle et al. (2008).

In Section 2.6.2 it was mentioned that the use of quasi-likelihood for dependent data in the work of Niaparast (2009) is the same as using generalised estimating equations for a marginal model in which the mean and variance have been derived from the conditional model. There are several other links between the various estimation and design methods which we discuss here.

MQL is similar to a quasi-likelihood/GEE approach, but with an approximated marginal mean and variance. The approximation to the marginal mean is obtained by ignoring the random effects, and the variance approximation arises from a first order Taylor series expansion (Breslow and Clayton, 1993). The observation that the marginal mean approximation could be substantially improved by using the attenuation formula (3.20) led to the proposal of the adjusted MQL approximation, which resulted in designs much closer to the 'correct' answer derived using the MLNI approach in Section 3.4.

In some sense there is a relationship between the GEE-type methods and the MLNI approach for the logistic random intercept model when there are 2 points per block, because in this case the first two moments specify the entire probability distribution. An area for future work is to see whether it is possible to use a method analogous to that of Niaparast (2009), but instead evaluating the marginal mean and variance of the conditional model computationally. This would allow the computations behind MLNI to be used to calculate designs with more points per block, although these would be quasi-likelihood, rather than maximum likelihood designs.

8.2 Other models

In Chapters 4 and 5 we developed techniques for optimal designs in single and multiple dosing bioassays with individual variation. In the multiple dosing case, the designs are found within a restricted class which we do not claim is optimal overall. However with the new technique these designs are relatively inexpensive to compute and may be useful as a benchmark against which to measure arbitrary candidate designs. An obvious question is whether additional insights could assist in finding the overall optimal designs. Another is whether there are techniques to handle more complicated random effects structures.

Additional research into the benefits of using the multiple-dosing approach could be helpful for practitioners: at the moment these kinds of studies do not seem to be used in real applications. This may perhaps be related to the lack of software implementations for fitting the 'GLMM-plus-stopping-rule' model. However, such a package certainly seems like a realistic possibility.

The development of HGLMs in Lee and Nelder (2001) allows for the modelling of the dependence of dispersion components upon the covariates. This flexibility may be important in the robust product design setting, where the aim is to find values of the x_i such that the variance of the response is low, as exemplified in Lee, Nelder and Park (2011). Exploration of the impact on the choice of design in this scenario would be interesting.

8.3 General comments

In all of the areas studied in this thesis, the adoption of the techniques in a practical environment would be aided by the implementation of user-friendly software. The most useful tool would perhaps be a graphical platform for the comparison between given proposed designs and those found using the algorithms outlined here. Applied case studies would no doubt prove effective in spurring the most needed developments in the methodology.

The general direction of this thesis is extending the availability of variance-optimal designs to a broader range of more complex statistical models. The use of these models involves making further parametric assumptions about the data generating process. An interesting and important topic is the robustness of these analyses, together with the optimal designs, to systematic departures from the model assumptions. An ideal treatment would not specify a parametric alternative for the truth, instead confining the possible discrepancies to a set which defines a 'neighbourhood' of the approximately correct model. Li and Wiens (2011) conduct research in this direction for misspecified dose-response models.

References

Abdelbasit, K. M. and Plackett, R. L. (1983), 'Experimental design for binary data', *Journal of the American Statistical Association* **78**, 90–98.

Apostol, T. (1974), *Mathematical analysis*, Vol. 2, Addison-Wesley, Reading, MA.

Atkinson, A. C. (2008), 'Examples of the use of an equivalence theorem in constructing optimum experimental designs for random-effects nonlinear regression models', *Journal of Statistical Planning and Inference* **138**, 2595–2606.

Atkinson, A. C., Chaloner, K., Herzberg, A. M. and Juritz, J. (1993), 'Optimum experimental designs for properties of a compartmental model', *Biometrics* **49**, 325–337.

Atkinson, A. C., Donev, A. N. and Tobias, R. D. (2007), *Optimum experimental designs, with SAS*, Oxford University Press.

Atkinson, A. C. and Haines, L. M. (1996), Designs for nonlinear and generalized linear models, *in* S. Ghosh and C. R. Rao, eds, 'Handbook of Statistics', Vol. 13, Elsevier, Amsterdam.

Barndorff-Nielsen, O. (1983), 'On a formula for the distribution of the maximum likelihood estimator', *Biometrika* **70**, 343–365.

Bazzoli, C., Retout, S. and Mentré, F. (2010), 'Design evaluation and optimisation in multiple response nonlinear mixed effect models: PFIM 3.0', *Computer Methods and Programs in Biomedicine* **98**, 55–65.

Berger, M. P. F. and Tan, F. E. S. (2004), 'Robust designs for linear mixed effects models', *Journal of the Royal Statistical Society, Series C* **53**, 569–581.

Biedermann, S., Dette, H. and Zhu, W. (2006), 'Optimal designs for dose-response models with restricted design spaces', *Journal of the American Statistical Association* **101**, 747–759.

Billingsley, P. (2012), *Probability and measure*, Wiley, Hoboken, NJ.

Box, G. E. P. and Draper, N. R. (1959), 'A basis for the selection of a response surface design', *Journal of the American Statistical Association* **54**, 622–654.

Breslow, N. E. and Clayton, D. G. (1993), 'Approximate inference in generalized linear mixed models', *Journal of the American Statistical Association* **88**, 9–25.

Broström, G. (2011), *glmmML: Generalized linear models with clustering.* R package version 0.82-1.
URL: *http://CRAN.R-project.org/package=glmmML*

Burr, D. (1988), 'On errors-in-variables in binary regression–Berkson case', *Journal of the American Statistical Association* **83**, 739–743.

Cassity, C. R. (1965), 'Abscissas, coefficients, and error term for the generalized Gauss-Laguerre quadrature', *Mathematics of Computation* **19**, 287–296.

Chaganty, N. R. and Joe, H. (2004), 'Efficiency of generalized estimating equations for binary responses', *Journal of the Royal Statistical Society, Series B* **66**, 851–860.

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**, 273–304.

Cheng, C. S. (1995), 'Optimal regression designs under random block-effects models', *Statistica Sinica* **5**, 485–497.

Condra, L. W. (1993), *Reliability improvement with design of experiments*, Marcel Dekker, New York.

Cox, D. R. (1988), 'A note on design when response has an exponential family distribution', *Biometrika* **75**, 161–164.

Davison, A. C. (2003), *Statistical models*, Cambridge University Press.

Demidenko, E. (2004), *Mixed Models: Theory and Applications*, Wiley, Hoboken, NJ.

Dennis, J. E. and Schnabel, R. B. (1987), *Numerical methods for unconstrained optimization and nonlinear equations*, Society for Industrial and Applied Mathematics, Philadelphia, PA.

Dette, H. (1997), 'Designing experiments with respect to 'standardized' optimality criteria', *Journal of the Royal Statistical Society, Series B* **59**, 97–110.

Donev, A. N. (2004), 'Design of experiments in the presence of errors in factor levels', *Journal of Statistical Planning and Inference* **126**, 569–585.

Draper, N. R. and Smith, H. (1998), *Applied regression analysis*, 3rd edn, Wiley, Hoboken, NJ.

Dror, H. A. and Steinberg, D. M. (2006), 'Robust experimental design for multivariate generalized linear models', *Technometrics* **48**, 520–529.

Dror, H. A. and Steinberg, D. M. (2008), 'Sequential experimental designs for generalized linear models', *Journal of the American Statistical Association* **103**, 288–298.

Fedorov, V. V. and Hackl, P. (1997), *Model-oriented design of experiments*, Springer, New York.

Firth, D. and Hinde, J. P. (1997), 'On Bayesian D-optimum design criteria and the equivalence theorem in non-linear models', *Journal of the Royal Statistical Society, Series B* **59**, 793–797.

Ford, I., Torsney, B. and Wu, C. F. J. (1992), 'The use of a canonical form in the construction of locally optimal designs for non-linear problems', *Journal of the Royal Statistical Society, Series B* **54**, 569–583.

Furrer, R., Nychka, D. and Sain, S. (2010), *fields: Tools for spatial data.* R package version 6.3. **URL:** *http://CRAN.R-project.org/package=fields*

Gagnon, R. and Leonov, S. (2004), 'Optimal population designs for PK models with serial sampling', *Journal of Biopharmaceutical Statistics* **15**, 143–163.

Goldstein, H. and Rasbash, J. (1996), 'Improved approximations for multilevel models with binary responses', *Journal of the Royal Statistical Society, Series A* **159**, 505–513.

Goos, P. and Vandebroek, M. (2001), 'D-optimal response surface designs in the presence of random block effects', *Computational Statistics & Data Analysis* **37**, 433–453.

Gotwalt, C. M., Jones, B. A. and Steinberg, D. M. (2009), 'Fast computation of designs robust to parameter uncertainty for nonlinear settings', *Technometrics* **51**, 88–95.

Hadfield, J. D. (2010), 'MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package', *Journal of Statistical Software* **33**, 1–22.

Hamada, M. and Nelder, J. A. (1997), 'Generalized linear models for quality-improvernent experiments', *Journal of Quality Technology* **29**, 292–304.

Hougaard, P. (1995), 'Frailty models for survival data', *Lifetime data analysis* **1**, 255–273.

Jones, B. and Goos, P. (2007), 'A candidate-set-free algorithm for generating D-optimal split-plot designs', *Journal of the Royal Statistical Society, Series C* **56**, 347–364.

Jones, B. and Goos, P. (2009), 'D-optimal design of split-split-plot experiments', *Biometrika* **96**, 67–82.

Jones, B. and Nachtsheim, C. J. (2009), 'Split-plot designs: What, why, and how', *Journal of Quality Technology* **41**, 340–361.

Khuri, A. I., Mukherjee, B., Sinha, B. K. and Ghosh, M. (2006), 'Design issues for generalized linear models: A review', *Statistical Science* **21**, 376–399.

King, J. and Wong, W. K. (2000), 'Minimax D-optimal designs for the logistic model', *Biometrics* **56**, 1263–1267.

Konstantinou, M., Biedermann, S. and Kimber, A. (2011), Optimal designs for two-parameter nonlinear models with application to survival models, Technical Report (Methodology) M11/08, Southampton Statistical Research Institute, University of Southampton.

Lee, Y. and Nelder, J. A. (1996), 'Hierarchical generalized linear models', *Journal of the Royal Statistical Society, Series B* **58**, 619–678.

Lee, Y. and Nelder, J. A. (2001), 'Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions', *Biometrika* **88**, 987–1006.

Lee, Y. and Nelder, J. A. (2009), 'Likelihood inference for models with unobservables: another view', *Statistical Science* **24**, 255–269.

Lee, Y., Nelder, J. A. and Noh, M. (2007), 'H-likelihood: problems and solutions', *Statistics and Computing* **17**, 49–55.

Lee, Y., Nelder, J. A. and Park, H. (2011), 'HGLMs for quality improvement', *Applied Stochastic Models in Business and Industry* **27**, 315–328.

Lee, Y., Nelder, J. A. and Pawitan, Y. (2006), *Generalized Linear Models with Random Effects: Unified Analysis via h-likelihood*, Chapman and Hall/CRC, Boca Raton, FL.

Lele, S. R., Nadeem, K. and Schmuland, B. (2010), 'Estimability and likelihood inference for generalized linear mixed models using data cloning', *Journal of the American Statistical Association* **105**, 1617–1625.

Li, P. and Wiens, D. P. (2011), 'Robustness of design in dose–response studies', *Journal of the Royal Statistical Society, Series B* **73**, 215–238.

Liang, K. Y. and Zeger, S. L. (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**, 13–22.

Loeza-Serrano, S. and Donev, A. N. (2012), 'Construction of experimental designs for estimating variance components (in press)', *Computational Statistics & Data Analysis* .
URL: *http://dx.doi.org/10.1016/j.csda.2012.10.008*

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, 2nd edn, Chapman and Hall, Boca Raton, FL.

McCulloch, C. E. (1997), 'Maximum likelihood algorithms for generalized linear mixed models', *Journal of the American Statistical Association* **92**, 162–170.

McCulloch, C. E. and Searle, S. R. (2001), *Generalized, Linear, and Mixed Models*, Wiley-Interscience, Hoboken, NJ.

Melas, V. (2005), 'On the functional approach to optimal designs for nonlinear models', *Journal of Statistical Planning and Inference* **132**, 93–116.

Mentré, F., Mallet, A. and Baccar, D. (1997), 'Optimal design in random-effects regression models', *Biometrika* **84**, 429–442.

Meyer, R. K. and Nachtsheim, C. J. (1995), 'The coordinate-exchange algorithm for constructing exact optimal experimental designs', *Technometrics* **37**, 60–69.

Mielke, T. and Schwabe, R. (2010), Some considerations on the Fisher information in non-linear mixed effects models, *in* A. Giovagnoli, A. C. Atkinson and B. Torsney, eds, 'mODa 9–Advances in Model-Oriented Design and Analysis', Springer, New York, pp. 129–136.

Minkin, S. (1993), 'Experimental design for clonogenic assays in chemotherapy', *Journal of the American Statistical Association* **88**, 410–420.

Moerbeek, M. and Maas, C. J. M. (2005), 'Optimal experimental designs for multilevel logistic models with two binary predictors', *Communications in Statistics-Theory and Methods* **34**, 1151–1167.

Moerbeek, M., Van Breukelen, G. J. P. and Berger, M. P. F. (2001), 'Optimal experimental designs for multilevel logistic models', *Journal of the Royal Statistical Society, Series D* **50**, 17–30.

Molas, M. and Lesaffre, E. (2011), 'Hierarchical generalized linear models: The R package HGLMMM', *Journal of Statistical Software* **39**, 1–20.

Myers, R. H., Montgomery, D. C. and Anderson-Cook, C. M. (2009), *Response surface methodology: process and product optimization using designed experiments*, 3rd edn, Wiley, Hoboken, NJ.

Nelder, J. A. and Mead, R. (1965), 'A simplex method for function minimization', *The Computer Journal* **7**, 308–313.

Niaparast, M. (2009), 'On optimal design for a Poisson regression model with random intercept', *Statistics & Probability Letters* **79**, 741–747.

Niaparast, M. and Schwabe, R. (2013), 'Optimal design for quasi-likelihood estimation in Poisson regression with random coefficients', *Journal of Statistical Planning and Inference* **143**, 296–306.

Nocedal, J. and Wright, S. J. (1999), *Numerical Optimization*, Springer, New York.

Ogungbenro, K. and Aarons, L. (2011), 'Population Fisher information matrix and optimal design of discrete data responses in population pharmacodynamic experiments', *Journal of Pharmacokinetics and Pharmacodynamics* **38**, 449–469.

Ouwens, M. J. N. M., Tan, F. E. S. and Berger, M. P. F. (2002), 'Maximin D-optimal designs for longitudinal mixed effects models', *Biometrics* **58**, 735–741.

Ouwens, M. J. N. M., Tan, F. E. S. and Berger, M. P. F. (2006), 'A maximin criterion for the logistic random intercept model with covariates', *Journal of Statistical Planning and Inference* **136**, 962–981.

Overstall, A. M. and Forster, J. J. (2010), 'Default Bayesian model determination methods for generalised linear mixed models', *Computational Statistics & Data Analysis* **54**, 3269–3288.

Press, W. H., Teuklosky, S. A., Vetterling, W. T. and Flannery, B. P. (1992), *Numerical recipes in C*, 3rd edn, Cambridge University Press.

Pukelsheim, F. (1987), 'Information increasing orderings in experimental design theory', *International Statistical Review* **55**, 203–219.

Pukelsheim, F. and Rieder, S. (1992), 'Efficient rounding of approximate designs', *Biometrika* **79**, 763–770.

R Development Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org*

Retout, S., Comets, E., Samson, A. and Mentré, F. (2007), 'Design in nonlinear mixed effects models: Optimization using the Fedorov–Wynn algorithm and power of the Wald test for binary covariates', *Statistics in Medicine* **26**, 5162–5179.

Retout, S. and Mentré, F. (2003), 'Further developments of the Fisher information matrix in nonlinear mixed effects models with evaluation in population pharmacokinetics', *Journal of Biopharmaceutical Statistics* **13**, 209–227.

Ridout, M. S. and Fenlon, J. S. (1991), ‘Analysing dose-mortality data when doses are subject to error’, *Annals of Applied Biology* **119**, 191–201.

Ridout, M. S., Fenlon, J. S. and Hughes, P. R. (1993), ‘A generalized one-hit model for bioassays of insect viruses’, *Biometrics* **49**, 1136–1141.

Robinson, T. J., Myers, R. H. and Montgomery, D. C. (2004), ‘Analysis considerations in industrial split-plot experiments with non-normal responses’, *Journal of Quality Technology* **36**, 180–192.

Robinson, T. J., Wulff, S. S., Montgomery, D. C. and Khuri, A. I. (2006), ‘Robust parameter design using generalized linear mixed models’, *Journal of Quality Technology* **38**, 65–75.

Rodriguez, G. and Goldman, N. (1995), ‘An assessment of estimation procedures for multilevel models with binary responses’, *Journal of the Royal Statistical Society, Series A* **158**, 77–89.

Ronnegard, L., Shen, X. and Moudud, A. (2010), ‘hglm: A package for fitting hierarchical generalized linear models’, *The R Journal* **2**, 20–28.
URL: *http://journal.r-project.org/archive/2010-2/RJournal_2010-2_Roennegaard~et~al.pdf*

Russell, K. G., Eccleston, J. A., Lewis, S. M. and Woods, D. C. (2009), ‘Design considerations for small experiments and simple logistic regression’, *Journal of Statistical Computation and Simulation* **79**, 81–91.

Russell, K. G., Woods, D. C., Lewis, S. M. and Eccleston, J. A. (2009), ‘D-optimal designs for Poisson regression models’, *Statistica Sinica* **19**, 721–730.

Silvey, S. D. (1980), *Optimal design: an introduction to the theory for parameter estimation*, Chapman and Hall, Boca Raton, FL.

Sinha, S. K. and Xu, X. (2011), ‘Sequential optimal designs for generalized linear mixed models’, *Journal of Statistical Planning and Inference* **141**, 1394–1402.

Smits, P. H. and Vlak, J. M. (1988), ‘Biological activity of spodoptera exigua nulear polyhedrosis virus against s. exigua larvae.’, *Journal of Invertebrate Pathology* **52**, 107–114.

Stewart, G. W. (1980), ‘The efficient generation of random orthogonal matrices with an application to condition estimation’, *SIAM Journal on Numerical Analysis* **17**, 403–409.

Stufken, J. and Yang, M. (2012), ‘On locally optimal designs for generalized linear models with group effects’, *Statistica Sinica* **22**, 1765–1786.

Tang, P. K. and Bacon-Shone, J. (1992), Bayesian optimal designs for probit regression with errors-in-variables, Technical report, University of Hong Kong. Dept. of Statistics.

Tekle, F. B., Tan, F. E. S. and Berger, M. P. F. (2008), ‘Maximin D-optimal designs for binary longitudinal responses’, *Computational Statistics & Data Analysis* **52**, 5253–5262.

Tsutakawa, R. K. (1972), ‘Design of experiment for bioassay’, *Journal of the American Statistical Association* **67**, 584–590.

Turner, H. and Firth, D. (2010), *Bradley-Terry models in R: The BradleyTerry2 package*. R package version 0.9-4.
URL: *http://CRAN.R-project.org/package=BradleyTerry*

Waite, T. W., Woods, D. C. and Waterhouse, T. H. (2012), Designs for generalized linear models with random block effects, Technical Report (Methodology) M12/02, Southampton Statistical Sciences Research Institute, University of Southampton.

Wedderburn, R. W. M. (1974), 'Quasi-likelihood functions, generalized linear models and the Gauss-Newton method', *Biometrika* **61**, 439–447.

Wiens, D. P. (1992), 'Minimax designs for approximately linear regression', *Journal of Statistical Planning and Inference* **31**, 353–371.

Williams, D. (1991), *Probability with martingales*, Cambridge University Press.

Woods, D. C., Lewis, S. M., Eccleston, J. A. and Russell, K. G. (2006), 'Designs for generalized linear models with several variables and model uncertainty', *Technometrics* **48**, 284–292.

Woods, D. C. and Van de Ven, P. (2011), 'Block designs for experiments with correlated non-normal response', *Technometrics* **53**, 173–182.

Xue, X. and Brookmeyer, R. (1997), 'Regression analysis of discrete time survival data under heterogeneity', *Statistics in Medicine* **16**, 1983–1993.

Yang, J., Mandal, A. and Majumdar, D. (2012), 'Optimal designs for two-level factorial experiments with binary response', *Statistica Sinica* **22**, 885–907.

Yang, M. (2010), 'On the de la Garza Phenomenon', *The Annals of Statistics* **38**, 2499–2524.

Yang, M. and Stufken, J. (2009), 'Support points of locally optimal designs for nonlinear models with two parameters', *The Annals of Statistics* **37**, 518–541.

Yang, M., Zhang, B. and Huang, S. (2011), 'Optimal designs for generalized linear models with multiple design variables', *Statistica Sinica* **21**, 1415–1430.

Zeger, S. L. and Karim, M. R. (1991), 'Generalized linear models with random effects: a Gibbs sampling approach', *Journal of the American Statistical Association* **86**, 79–86.

Zeger, S. L., Liang, K. Y. and Albert, P. S. (1988), 'Models for longitudinal data: a generalized estimating equation approach', *Biometrics* **44**, 1049–1060.