Journal of Statistical Planning and Inference 144 (2014) 152–159

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Optimal blocked minimum-support designs for non-linear models

P.M. van de Ven [a], D.C. Woods [b],*

[a] Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands
[b] Southampton Statistical Sciences Research Institute, University of Southampton, Southampton SO17 1BJ, UK

ARTICLE INFO

Available online 21 February 2013

Keywords:
Binary data
Block designs
Count data
D-optimality
Generalized estimating equations
Generalized linear models
Michaelis–Menten model

ABSTRACT

Finding optimal designs for experiments for non-linear models and dependent data is a challenging task. We show how the problem simplifies when the search is restricted to designs that are minimally supported; that is, the number of distinct runs (treatments) is equal to the number of unknown parameters, p, in the model. Under this restriction, the problem of finding a locally or pseudo-Bayesian D-optimal design decomposes into two simpler problems that are more widely studied. The first is that of finding a minimum-support D-optimal design d_1 with p runs for the corresponding model for the mean but assuming independent observations. The second problem is finding a D-optimal block design for assigning the treatments in d_1 to the experimental units. We find and assess optimal minimum-support designs for three examples, each assuming a mean model from a different member of the exponential family: binomial, Poisson and normal. In each case, the efficiencies of the designs are compared to the optimal design where the restriction on the number of distinct support points is relaxed. The optimal minimum-support designs are found to often perform satisfactorily under both local and Bayesian D-optimality for concentrated prior distributions. The results are also relatively insensitive to the assumed degree of dependence in the data.

1. Introduction

Many experiments aim to model a non-linear relationship between a response and several explanatory variables. If a binary or count response variable is observed, an appropriate generalized linear model (GLM) can be assumed to describe this relationship. In other cases, the outcome may be continuous and normally distributed, but the relationship between the mean and the explanatory variables may be non-linear. The data observed from different runs in the experiment are often assumed to be independent and the parameters in the model are estimated using maximum likelihood estimation techniques. It is, however, not uncommon for experiments to be performed in blocks; that is, different runs are performed, for example, on different days, by different scientists or operators, or using different batches of material. Such situations may induce dependence between the observations within a block, whilst observations in different blocks remain independent. An estimation method that takes into account this dependence structure increases the accuracy of inferences made from the experimental data. The blocking structure should also be taken into account when designing the experiment.

* Corresponding author. Tel.: +44 02380 595117; fax: +44 2380 595147.
E-mail addresses: p.vandeven@vumc.nl (P.M. van de Ven), D.Woods@southampton.ac.uk (D.C. Woods).

0378-3758/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2013.02.001

Most literature on design for non-linear models and dependent data has been concerned with non-linear mixed effect models, see the seminal paper of Mentré et al. (1997), typically in a clinical or biological setting (see also Han and Chaloner, 2004; Atkinson, 2008, and references therein). Such methodology assumes a conditional, or subject-specific, modeling approach, with potentially differing model parameters for each block. In this paper, we find designs for marginal, or population-averaged, models where the dependencies in the data do not arise from subject-specific parameters; see also Hughes-Oliver (1998) and Atkinson and Ucinski (2004). For the linear model, where there has been substantial work on designs with random block effects (see, for example, Goos and Vandebroek, 2001), these two modeling paradigms coincide.

Recently, the first general results and methods for finding optimal blocked designs for discrete data have been published. Niaparast (2009) presented local *D*-optimal designs for Poisson regression models with a random intercept in the linear predictor and a single explanatory variable. Woods and van de Ven (2011) presented approaches for finding *D*-optimal designs for non-normal data and generalized estimating equations (GEE models) and any number of explanatory variables.

Designs for non-linear models generally suffer from their performance depending on the values of the unknown model parameters. Hence, either an initial guess of these parameters is required (Chernoff, 1953), perhaps as part of a sequential strategy (Dror and Steinberg, 2008), or a Bayesian, or pseudo-Bayesian, design is required; see Han and Chaloner (2004), Woods et al. (2006) and Section 2 of this paper.

The focus in this paper is on designs that are minimally supported, i.e. the number of distinct design points, or treatments, is equal to the number of unknown parameters in the model; see also Cheng (1995). We present methodology for finding minimum-support block designs to estimate the model parameters in a non-linear model with dependent observations for both continuous and discrete responses. The restriction to minimally supported designs allows a decomposition of the pseudo-Bayesian *D*-optimality objective function, leading to two, simpler, optimization problems (Section 3). For some experiments, there may also be practical advantages in reducing the number of treatments employed, such as reduced cost; for example, when the experiment relies on the construction of templates or formulations in manufacturing or chemistry. The decomposition of the objective function allows either the analytic derivation of optimal designs or the computational complexity of finding optimal designs to be reduced and we show, through a series of examples (Section 4), that minimum-support designs can suffer only a minor loss in performance compared to unrestricted designs.

2. Models and design criteria

Consider a continuous or discrete response $Y(\mathbf{x})$ that may depend on the values taken by m explanatory variables $\mathbf{x}^{\mathrm{T}}=(x_1,x_2,\ldots,x_m)$. In an experiment, responses are observed for different settings of the explanatory variables according to a design d consisting of N experimental runs. In run j, the jth experimental unit receives treatment $\mathbf{x}_j^{\mathrm{T}}=(x_{1j},\ldots,x_{mj})$, chosen from a bounded design space $\mathcal{X}\subset\mathbb{R}^m$ $(j=1,\ldots,N)$. We assume that the units are arranged in b blocks of size k_l $(l=1,\ldots,b)$ so that $k_1+k_2+\cdots+k_b=N$. The entries in the design $d=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\}$ and the observations $\mathbf{Y}=(Y(\mathbf{x}_1),Y(\mathbf{x}_2),\ldots,Y(\mathbf{x}_N))^{\mathrm{T}}$ are ordered by block and by unit within block. All pairs of observations made in different blocks are assumed independent but observations within the same block may be dependent.

We are interested in finding efficient designs for estimating the unknown parameters $\boldsymbol{\beta}=(\beta_1,\beta_2,\ldots,\beta_p)^{\mathrm{T}}$ in a marginal model for the mean response $\mathrm{E}[Y(\mathbf{x})]=\mu(\mathbf{x},\boldsymbol{\beta})$. We consider a general class of models for which the inverse of the model-based asymptotic variance–covariance matrix for an estimator $\widehat{\boldsymbol{\beta}}$ is of the form

$$\mathbf{M}=\mathbf{F}^{\mathrm{T}}\mathbf{V}^{-1/2}\mathbf{R}^{-1}\mathbf{V}^{-1/2}\mathbf{F}, \tag{1}$$

where only the matrices $\mathbf{F}$ and $\mathbf{V}$ depend on the design d and the parameter values $\boldsymbol{\beta}$. The matrix $\mathbf{R}$ is a block-diagonal correlation matrix. Common examples include marginal models for correlated discrete data estimated using the GEE approach, under some assumptions, and non-linear models with additive correlated normally distributed errors.

Under a GEE model for a discrete response, the mean and variance of the observations are assumed to come from an appropriate GLM such that $\mathrm{Var}[Y(\mathbf{x})]=v[\mu(\mathbf{x},\boldsymbol{\beta})]/\phi$, where ϕ is a constant scale parameter and $v(\cdot)$ is the variance function of the GLM (Liang and Zeger, 1986). The mean is related to $\mathbf{x}$ through $g[\mu(\mathbf{x},\boldsymbol{\beta})]=\mathbf{f}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}$, where $g(\cdot)$ is the link function of the GLM and the product $\mathbf{f}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}$ is the linear predictor, with the p-vector $\mathbf{f}(\mathbf{x})$ holding known functions of $\mathbf{x}$. The dependence in the data is modeled by means of a "working correlation" matrix $\mathbf{R}$, which is assumed known. This matrix will typically have a standard structure and is not necessarily equal to the actual correlation structure in $\mathbf{Y}$. The model-based estimator for the asymptotic variance–covariance matrix for the GEE estimator $\widehat{\boldsymbol{\beta}}$ is given by

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}})=(\mathbf{X}^{\mathrm{T}}\boldsymbol{\Delta}\mathbf{V}^{-1/2}\mathbf{R}^{-1}\mathbf{V}^{-1/2}\boldsymbol{\Delta}\mathbf{X})^{-1}, \tag{2}$$

where $\mathbf{X}$ is the $N\times p$ model matrix with rows $\mathbf{f}^{\mathrm{T}}(\mathbf{x}_j)$, $\boldsymbol{\Delta}=\mathrm{diag}\{1/g'[\mu(\mathbf{x}_j,\boldsymbol{\beta})]\}$ and $\mathbf{V}=\mathrm{diag}\{\mathrm{Var}[Y(\mathbf{x}_j)]\}$; see Lee et al. (2006, p. 75). The inverse of the variance–covariance matrix given in (2) is of the form specified in (1) with $\mathbf{F}=\boldsymbol{\Delta}\mathbf{X}$. This model-based estimator is derived under the assumption that the working correlation is exactly equal to the true correlation structure, which may not hold. However, the choice of the design has been shown to be relatively insensitive to the exact correlation structure (see Woods and van de Ven, 2011).