Designs for generalized linear models with random block effects via information matrix approximations

T. W. Waite[*] and D. C. Woods[†]

Southampton Statistical Sciences Research Institute,
University of Southampton, SO17 1BJ, U.K.

Abstract

The selection of optimal designs for generalized linear mixed models is complicated by the fact that the Fisher information matrix, on which most optimality criteria depend, is computationally expensive to evaluate. Our focus is on the design of experiments for likelihood estimation of parameters in the conditional model. We provide two novel approximations that substantially reduce the computational cost of evaluating the information matrix by complete enumeration of response outcomes, or Monte Carlo approximations thereof: (i) an asymptotic approximation which is accurate when there is strong dependence between observations in the same block; (ii) an approximation via Kriging interpolators. For logistic random intercept models, we show how interpolation can be especially effective for finding pseudo-Bayesian designs that incorporate uncertainty in the values of the model parameters. The new results are used to provide the first evaluation of the efficiency, for estimating conditional models, of optimal designs from closed-form approximations to the information matrix derived from marginal models. It is found that correcting for the marginal attenuation of parameters in binary-response models yields much improved designs, typically with very high efficiencies. However, in some experiments exhibiting strong dependence, designs for marginal models may still be inefficient for conditional modelling. Our asymptotic results provide some theoretical insights into why such inefficiencies occur.

Key words: Bayesian design; Binary response; Blocked experiment; Count response; Generalized linear mixed model; Kriging; Outcome-enumeration; Quasi-likelihood.

1 Introduction

There is increasing recognition of the need to design experiments in situations where a linear model with only fixed effects cannot adequately capture the essential features of the data. In particular, there is a growing body of work on optimal design for generalized linear models (for example, Chaloner & Larntz 1989, Woods et al. 2006, Yang et al. 2011) which can be used when the response variable follows a non-normal distribution from the exponential family. An even more substantial literature addresses the problem of optimal design when there is heterogeneity between blocks in an experiment, using a linear mixed model for normally distributed responses with random block effects (for example, Cheng 1995 and Goos & Vandebroek 2001). In many practical contexts, such as the industrial experiment described by Woods & Van de Ven (2011), both of these features (non-normality and heterogeneity) are present. For such experiments, particularly where the response is discrete, for example binary or count data, generalized linear mixed models may be an appropriate modelling choice. In this paper, we develop and compare optimal design methodologies for this family of models.

We find D-optimal designs, that is, designs that minimize the volume of the asymptotic confidence ellipsoid for the model parameters by maximizing the determinant of the Fisher information matrix. Dependence of an optimal design on the unknown values of the parameters is addressed using a pseudo-Bayesian approach. The key technical difficulty in the construction of D-optimal designs for generalized linear mixed models is that the information matrix is computationally expensive to evaluate.

The adoption of a mixed model implies a marginal distribution for the response with intra-block correlation: two responses from units within the same block are correlated and responses from units in

---

[*]t.w.waite@southampton.ac.uk

[†]d.woods@southampton.ac.uk

different blocks are uncorrelated. Existing approaches to optimal design for correlated discrete responses (see, for example, Moerbeek & Maas 2005, Niaparast 2009 and Woods & Van de Ven 2011) are tailored for inferential methods, such as quasi-likelihood and generalized estimating equations, that use only the first and second order moments of the marginal distribution, and approximations thereof, for parameter estimation. Our focus is on design for direct (likelihood) estimation of parameters in the conditional model for which we present novel asymptotic and computational approximations to the information matrix. For binary data, we also adapt and extend marginal approximations to provide more efficient designs for the conditional model. We compare designs from the various approximations to those found from computationally expensive "gold standard" approximations throughout our examples, using either naïve outcome-enumeration or Monte Carlo methods (Section 2.2).

2 Preliminaries

2.1 Generalized linear mixed models for blocked experiments

We denote the response for the jth unit in the ith block by y_{ij}, and the corresponding treatment vector of values taken by the q controllable variables by $x_{ij} \in \mathcal{X} \subseteq \mathbb{R}^q$ ($i = 1, \ldots, n$, $j = 1, \ldots, m_i$). Also let $\zeta_i = (x_{i1}, \ldots, x_{im_i}) \in \mathcal{X}^{m_i}$ denote the m_i treatment vectors in the ith block. Then, for a generalized linear mixed model, there is a vector, u_i, of r random effects associated with the ith block in the experiment. Conditional on u_i, the responses in block i are independent and follow an exponential family distribution, $y_{ij}|u_i \sim \pi(x_{ij}; u_i, \beta)$, with mean $\mu_{ij} = \mu(x_{ij}; u_i, \beta)$ and variance $\varphi v(x_{ij}; u_i, \beta)$. For the models we consider, the dispersion parameter $\varphi = 1$. The mean function $\mu(x; u, \beta)$ is defined by

$$g\{\mu(x; u, \beta)\} = \nu(x; u, \beta), \quad \nu(x; u, \beta) = f^{\mathrm{T}}(x)\beta + z^{\mathrm{T}}(x)u, \tag{1}$$

where $f : \mathcal{X} \to \mathbb{R}^p$ and $z : \mathcal{X} \to \mathbb{R}^r$ are known vectors of regressor functions, with z typically a subvector of f, and β is a p-vector of fixed regression parameters. The link function g relates the linear predictor ν to the mean response. Denote by h the inverse link function, g^{-1}. To fully determine the model, assumptions about the distribution of u_i are necessary; we specify independent $u_i \sim \mathrm{MVN}(0, G)$, with G an arbitrary covariance matrix. The presence of random effects in the linear predictor introduces a correlation between observations from experimental units in the same block. In this paper, we focus principally on random intercept models appropriate for blocked experiments, where $r = 1$, $G = (\sigma^2)$, $\sigma^2 > 0$, and $z(x) = (1)$.

We assume for simplicity that $m_i = m$, and say treatment blocks $\zeta_1, \zeta_2 \in \mathcal{X}^m$ are *equivalent* if one can be obtained from the other by treatment permutation. Without loss of generality, we may assume that the blocks are ordered so that the design is supported on $\zeta_1, \ldots, \zeta_b$ ($1 \leq b \leq n$), and that no pair from among the first b blocks is equivalent. Let w_k, $k = 1, \ldots, b$, be the proportion of blocks equivalent to the kth support block ζ_k, then we have the following concise notation for a design:

$$\xi = \left\{ \begin{array}{ccc} \zeta_1, & \cdots, & \zeta_b \\ w_1, & \cdots, & w_b \end{array} \right\}, \tag{2}$$

where $0 < w_k \leq 1$ and $\sum_{k=1}^{b} w_k = 1$. As defined above, nw_k is a positive integer. We focus on approximate block designs, which relax this constraint (see also Cheng 1995). Note that we do not impose any restrictions on the form of ζ_k, and that designs may differ in the w_k, the ζ_k and the value of b. We restrict to designs with finite support, $b \leq n < \infty$.

2.2 Information matrix

Let θ denote the complete vector of parameters for model (1). Thus θ includes the fixed effects parameters β as well as any parameters specifying the distribution of u_i. Denote by M_β the information matrix for β, holding all other components of θ fixed. The use of M_β is appropriate for assessing the precision of a maximum likelihood estimator $\hat{\beta}$ assuming known variance components. In common with many papers on design for both linear mixed models (Cheng 1995, Goos & Vandebroek 2001) and specific examples of their generalized counterparts (Moerbeek & Maas 2005, Tekle et al. 2008, Niaparast & Schwabe 2013), we do not consider the additional variability in $\hat{\beta}$ introduced when the variance components also require estimation.

For the approximate block design ξ in (2), the information matrix M_β depends on θ and, as observations in different blocks are independent, can be decomposed into a weighted sum of the information matrices for each support block,

$$M_\beta(\xi, \theta) = \sum_{k=1}^{b} w_k M_\beta(\zeta_k, \theta) \,. \tag{3}$$

The information matrix for an arbitrary block $\zeta = (x_1, \ldots, x_m) \in \mathcal{X}^m$ is

$$M_\beta(\zeta, \theta) = F^{\mathrm{T}} E_Y \left\{ P(Y|\theta, \zeta)^{-2} \left(\frac{\partial P(Y|\theta, \zeta)}{\partial \eta} \right) \left(\frac{\partial P(Y|\theta, \zeta)}{\partial \eta} \right)^{\mathrm{T}} \right\} F \,, \tag{4}$$

where $Y = (y_1, \ldots, y_m)^{\mathrm{T}}$ denotes the response vector or outcome corresponding to ζ, $P(Y|\theta, \zeta)$ is the marginal likelihood of the model parameters, $\eta = (f^{\mathrm{T}}(x_1)\beta, \ldots, f^{\mathrm{T}}(x_m)\beta)^{\mathrm{T}}$, and $F = [f(x_1), \ldots, f(x_m)]^{\mathrm{T}}$ is the model matrix. The likelihood and its derivative are of the form

$$P(Y|\theta, \zeta) = \int_{\mathbb{R}^r} P(Y|u, \theta, \zeta) f_u(u) du \,, \quad \frac{\partial P(Y|\theta, \zeta)}{\partial \eta} = \int_{\mathbb{R}^r} \frac{\partial P(Y|u, \theta, \zeta)}{\partial \eta} f_u(u) du \,, \tag{5}$$

where $P(Y|u, \theta, \zeta)$ is the (exponential family) conditional probability density of Y given u and f_u is the density function of an MVN$(0, G)$ random variable. Typically a closed form for the partial derivative of the conditional density is available. For random intercept models, the integrals in (5) can be evaluated numerically using Gauss-Hermite quadrature.

For models with binary response, the expectation in (4) can be evaluated by enumeration of outcomes $Y \in \{0, 1\}^m$. Expanding the expectation, we obtain

$$M_\beta(\zeta, \theta) = F^{\mathrm{T}} \sum_{Y \in \{0,1\}^m} P(Y|\theta, \zeta)^{-1} \left(\frac{\partial P(Y|\theta, \zeta)}{\partial \eta} \right) \left(\frac{\partial P(Y|\theta, \zeta)}{\partial \eta} \right)^{\mathrm{T}} F \,, \tag{6}$$

where the sum is over all possible response patterns in block ζ. An obvious approximation to information matrix (4) is via (6) with numerical approximation of (5) using quadrature. We call this approach *naïve outcome-enumeration*. Clearly, for even moderately sized blocks, such an approximation will be computationally expensive.

For other response distributions, such as Poisson, the expectation in (4) can be approximated, in principle, by Monte Carlo sampling of response vectors Y. In practice, to obtain reasonable precision in the approximation of the information matrix using this method, it is necessary to consider many more than the 2^m possible distinct outcomes obtained from a binary model.

2.3 Optimality criteria

We study both locally D-optimal designs, i.e. $\xi_D^* = \arg\max_\xi |M_\beta(\xi, \theta)|$ for an assumed value of θ, and (pseudo-)Bayesian designs. From (3) and an application of Caratheodory's theorem (e.g. Silvey 1980, p.16), it follows that there is always a locally D-optimal design supported on at most $p(p+1)/2 + 1$ distinct blocks. The pseudo-Bayesian approach may be used to construct a design that is more robust to misspecification of the model parameters, and requires specification of a prior distribution, Λ, for θ. Given Λ, ξ is Bayesian D-optimal if it maximizes $\psi(\xi) = E_\theta\{\log |M_\beta(\xi, \theta)|\}$ (Chaloner & Larntz 1989). We do not assume that the resulting analysis will be Bayesian, or that it will use prior distribution Λ. Care must be taken when the prior distribution has unbounded support; see Waite (2013).

3 Approximations via marginal models

3.1 Marginal quasi-likelihood

Breslow & Clayton (1993) discussed marginal quasi-likelihood as a computationally inexpensive, approximate method for estimating the parameters of a generalized linear mixed model. The method is indirect in that it applies standard quasi-likelihood equations for dependent data (McCullagh & Nelder 1989,

Sec. 9·3) to a linearization of the model about the mean value of the random effects. An information matrix approximation corresponding to this method is

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

where F_k is the model matrix for ζ_k, $V_k = \mathcal{V}(\zeta_k,\theta)$ is determined from $\mathcal{V}(\zeta,\theta) = W(\zeta,\theta)^{-1} + Z(\zeta)GZ(\zeta)^{\mathrm{T}}$, $W(\zeta,\theta)$ is the diagonal matrix with entries $v(x_1;0,\beta),\ldots,v(x_m;0,\beta)$, and $Z(\zeta) = [z(x_1),\ldots,z(x_m)]^{\mathrm{T}}$. For design using similar methods, see Moerbeek & Maas (2005).

There are several higher-order marginal quasi-likelihood approximations in the literature, for example Goldstein & Rasbash (1996). An approximation to the information matrix using a second order method was derived in a 2012 University of Southampton PhD thesis by T. W. Waite. Use of this approximation does not result in better designs, so we omit the results here. The marginal quasi-likelihood approximation is similar to the first-order approximations used in the design of pharmacokinetic studies (see, for example, Retout & Mentré 2003).

3.2 Generalized estimating equations

Generalized estimating equations (Liang & Zeger 1986) may be used to estimate parameters when the marginal distribution of the response follows a generalized linear model, making use of a 'working correlation' matrix that need not be equal to the true correlation matrix. Typically, a standard structure is used for the working correlation, such as exchangeable, autoregressive or nearest neighbour. However these assumptions are incompatible with most known probability models for dependent discrete responses, in which the correlation is a nontrivial function of the treatments and parameters. Indeed there may not exist any probability model achieving these simple correlation structures with the required univariate marginal distributions if, for example, the working correlation violates the bounds on correlation for binary data (Joe 1997, Ch.7). Nonetheless, the estimators retain consistency under misspecification of the correlation structure and may still be highly efficient (Chaganty & Joe 2004). Note that here we use generalized estimating equations only to obtain an approximation to the mixed model information matrix.

Woods & Van de Ven (2011) found designs for marginal generalized linear models that are D-optimal for the generalized estimating equation method under the assumption that the true correlation structure corresponds to a specified working correlation structure. They also found that the resulting designs were robust to a general class of departures from this correlation assumption. Denote the parameters of the assumed marginal model by β^*, the correlation parameter by ρ, and assume the marginal model has the same link and variance functions as the conditional model. Then for exchangeable correlation, the inverse asymptotic covariance matrix is

$$M_\beta^{\mathrm{gen}}(\xi,\beta^*,\rho) = \sum_{k=1}^{b} w_k F_k^{\mathrm{T}} D_k \{(V_k^*)^{1/2} R(\rho)(V_k^*)^{1/2}\}^{-1} D_k F_k\,,$$

where D_k is the diagonal matrix with entries $1/g'(\mu_{k1}^*),\ldots,1/g'(\mu_{km}^*)$, $\mu_{kj}^* = h\{f^{\mathrm{T}}(x_{kj})\beta^*\}$, V_k^* is the diagonal matrix with entries $v(x_{k1};0,\beta^*),\ldots,v(x_{km};0,\beta^*)$, and $R(\rho) = (1-\rho)I_m + \rho 1_m 1_m^{\mathrm{T}}$, with I_m the $m \times m$ identity matrix and 1_m an m-vector of ones.

3.3 Binary response: adjustment for attenuation of parameters

Use of marginal quasi-likelihood for the logistic random effects model results in the assumption that the marginal mean has the form $E(y_{ij}) \approx g^{-1}\{f^{\mathrm{T}}(x_{ij})\beta\}$ (Breslow & Clayton 1993). Zeger et al. (1988) showed that a better approximation to the marginal mean is given by a logistic relationship with attenuated coefficients,

$$E(y_{ij}) \approx g^{-1}\left\{ f^{\mathrm{T}}(x_{ij})\beta / \sqrt{1 + c^2 z(x_{ij})^{\mathrm{T}} G z(x_{ij})} \right\}, \tag{7}$$

where $c = 16\sqrt{3}/(15\pi)$. For random intercept models this reduces to

$$E(y_{ij}) \approx g^{-1}\{f^{\mathrm{T}}(x_{ij})\beta_{\mathrm{att}}\}, \quad \beta_{\mathrm{att}} = \beta\,(1 + c^2\sigma^2)^{-1/2}. \tag{8}$$

4

This suggests that for the logistic random intercept model, more efficient designs might be obtained by adjusting the parameter values to better approximate the marginal mean using (8). Explicitly, we define the *adjusted marginal quasi-likelihood* information matrix by

$$M_\beta^{\mathrm{adj}}(\xi, \theta) = M_\beta^{\mathrm{marg}}(\xi, \theta_{\mathrm{adj}}), \quad \theta_{\mathrm{adj}} = \left(\beta_{\mathrm{att}}^{\mathrm{T}}, \sigma^2\right)^{\mathrm{T}},$$

where M_β^{marg} is the information matrix for β under marginal quasi-likelihood. In models other than the random intercept the attenuation factor depends on x, so a constant adjustment cannot be applied for every design point. However, one possibility for a similar approximation may be to apply quasi-likelihood or generalized estimating equations using (7) as the marginal mean.

To extend the methods of Woods & Van de Ven (2011), we also take account of parameter attenuation by forming the *adjusted generalized estimating equation* approximation,

$$M_\beta^{\mathrm{adj.\ gen.}}(\xi, \theta, \rho) = M_\beta^{\mathrm{gen.}}(\xi, \beta_{\mathrm{att}}, \rho). \tag{9}$$

Here we either choose a value of ρ following the guidelines laid out, for estimation, by Chaganty & Joe (2004), or treat ρ as a tuning parameter, i.e. we choose the value of ρ such that the corresponding D-optimal design using (9) maximizes $|M_\beta|$ approximated via naïve outcome-enumeration.

4 Theoretical and computational direct approximations for the logistic random intercept model

4.1 Asymptotic outcome-enumeration

For the logistic random intercept model, the important case of large σ^2 results in substantial block-to-block variability and poses a more difficult design problem. In this case, responses in the same block are strongly dependent, and the adjusted marginal quasi-likelihood and adjusted generalized estimating equation designs may perform quite poorly (see Section 5.2). Moreover, for large σ^2 the naïve outcome-enumeration approximation becomes even more computationally expensive, as more quadrature points are required to maintain accuracy in the approximation of the integrals in (5). In this section, we develop asymptotic, $\sigma^2 \to \infty$, expressions for $P(Y|\theta, \zeta)$ and its derivatives which are combined with (6) to provide a new, direct approximation to the information matrix for large finite σ^2. The additional approximation enables selection of efficient designs for large σ^2 at low computational cost compared to naïve outcome-enumeration (Section 5.2). Our main results are in Theorems 1–3; first we define some necessary assumptions and notation.

For fixed values of the conditional parameters, the 'marginal effects' in β_{att} attenuate to zero as $\sigma^2 \to \infty$. In order to approximate the more interesting and realistic case where both σ^2 is large and there are non-zero marginal effects, we assume the following asymptotic conditions.

Assumption 1. $\beta_{att} = \beta/\sqrt{1 + c^2\sigma^2}$ *is fixed.*

Assumption 2. *For each j, either $\eta_j^* = f^{\mathrm{T}}(x_j)\beta_{att}$ is fixed or there exists $l \neq j$ such that η_l^* is fixed and $\eta_l^* - \eta_j^* = o(\sigma^{-1})$.*

In order to meet these conditions, we allow the x_j to vary with σ^2. A simple asymptotic approximation to the information matrix could be derived by treating all η_j^* as distinct and fixed as $\sigma^2 \to \infty$. However, such an approximation would be very poor for designs with $\eta_l^* \approx \eta_j^*$ for some $l \neq j$. Our novel asymptotic framework allows consideration of the case where there is near-replication of linear predictor values in a block.

Assumptions 1 and 2 allow the partition of $\mathcal{S} = \{1, \ldots, m\}$ as $\mathcal{N}(j) \cup \mathcal{Z}(j) \cup \mathcal{P}(j)$ for each $j = 1, \ldots, m$, where $\mathcal{N}(j) = \{l : \eta_l - \eta_j \to -\infty\}$, $\mathcal{Z}(j) = \{l : \eta_l - \eta_j \to 0\}$, and $\mathcal{P}(j) = \{l : \eta_l - \eta_j \to \infty\}$. Intuitively, $\mathcal{N}(j)$, $\mathcal{Z}(j)$, $\mathcal{P}(j)$ are the respective sets of indices of linear predictors less than, similar to, and greater than η_j. The limiting expressions we develop for $P(Y|\theta, \zeta)$ and $\partial P(Y|\theta, \zeta)/\partial \eta_j$ depend on which elements of $\mathcal{S}$ belong to $\mathcal{N}$, $\mathcal{Z}$ and $\mathcal{P}$.

It will be useful to identify some particular classes of outcomes.

Definition 1. *Outcome $Y = (y_1, \ldots, y_m)^{\mathrm{T}}$ is increasing (within the block) if there exists $j' \in \mathcal{S}$ such that $y_l = 0$ when $\eta_l - \eta_{j'} < 0$ and $y_l = 1$ when $\eta_l - \eta_{j'} > 0$.*

5

Definition 2. *Outcome Y is* quasi-increasing *if there exists $j' \in \mathcal{S}$ such that $\mathcal{N}(j') \subseteq \mathcal{S}_0$ and $\mathcal{P}(j') \subseteq \mathcal{S}_1$, where $\mathcal{S}_0 = \{j : y_j = 0\}$ and $\mathcal{S}_1 = \{j : y_j = 1\}$, or, equivalently, if $\{\mathcal{S}_0 \cap \mathcal{P}(j')\} \cup \{\mathcal{S}_1 \cap \mathcal{N}(j')\} = \emptyset$.*

Any outcome that is increasing (with the same j') for all σ^2 is clearly also quasi-increasing.
We now make a further assumption necessary for our theorems.

Assumption 3. *There exists $A_j, B_j > 0$ such that $|\eta_l - \eta_j| > \sigma A_j$ for $l \in \{\mathcal{S}_0 \cap \mathcal{N}(j)\} \cup \{\mathcal{S}_1 \cap \mathcal{P}(j)\}$ and $|\eta_l - \eta_j| > \sigma B_j$ for all $l \in \{\mathcal{S}_1 \cap \mathcal{N}(j)\} \cup \{\mathcal{S}_0 \cap \mathcal{P}(j)\}$.*

This condition holds for large σ^2 by Assumptions 1 and 2; it implies that pairs of predictors which diverge asymptotically are at least $\min_{j=1,\dots,m}\{A_j, B_j\}\sigma$ apart.

Theorem 1 (Approximation of the likelihood). *Suppose that the outcome is quasi-increasing. Then there exists $j' \in \mathcal{S}$ such that $\{\mathcal{S}_0 \cap \mathcal{P}(j')\} \cup \{\mathcal{S}_1 \cap \mathcal{N}(j')\} = \emptyset$, and:*
(i) If $|\mathcal{S}_0 \cap \mathcal{Z}(j')| = 0$ or $|\mathcal{S}_1 \cap \mathcal{Z}(j')| = 0$, the outcome is increasing and, as $\sigma^2 \to \infty$,

$$P(Y|\theta, \zeta) = \max\left\{0, \Phi\left(-\max_{j \in \mathcal{S}_0}\{\eta_j/\sigma\}\right) - \Phi\left(-\min_{j \in \mathcal{S}_1}\{\eta_j/\sigma\}\right)\right\} + O(\sigma^{-1}). \tag{10}$$

(ii) If $|\mathcal{S}_0 \cap \mathcal{Z}(j')| \geq 1$ and $|\mathcal{S}_1 \cap \mathcal{Z}(j')| \geq 1$, then as $\sigma^2 \to \infty$,

$$P(Y|\theta, \zeta) = \frac{\phi(\eta_{j'}/\sigma)}{\sigma} \int_{-\infty}^{\infty} \{1 - h(t)\}^{|\mathcal{S}_0 \cap \mathcal{Z}(j')|} h(t)^{|\mathcal{S}_1 \cap \mathcal{Z}(j')|} dt$$
$$+ \sum_{l \in \mathcal{Z}(j')} O(\Delta_{lj'}/\sigma) + O(\sigma^{-2}), \tag{11}$$

where $\Delta_{lj} = \eta_l - \eta_j$. The integral has value 1 when $|\mathcal{Z}(j')| = 2$.

Theorem 2 (Approximation of derivatives). *(i) Assume $j \in \mathcal{S}$ is such that $\mathcal{N}(j) \subseteq \mathcal{S}_0$ and $\mathcal{P}(j) \subseteq \mathcal{S}_1$ which implies that the outcome is quasi-increasing. Then, for arbitrary $\epsilon > 0$,*

$$(2y_j - 1)\frac{\partial P(Y|\theta, \zeta)}{\partial \eta_j} = \frac{1}{\sigma}\phi\left(\frac{-\eta_j}{\sigma}\right)\left\{C_{I(j),J(j)}^{(1)} + \sum_{l \in \mathcal{S}_1 \cap \mathcal{Z}(j)\setminus\{j\}} \Delta_{lj} C_{I(j)-1,J(j)}^{(3)}\right.$$
$$\left. - \sum_{l \in \mathcal{S}_0 \cap \mathcal{Z}(j)\setminus\{j\}} \Delta_{lj} C_{I(j),J(j)-1}^{(3)}\right\} + \frac{1}{\sigma^2}\phi'\left(\frac{-\eta_j}{\sigma}\right)C_{I(j),J(j)}^{(2)}$$
$$+ \sum_{l \in \mathcal{Z}(j)} O(\Delta_{lj}^2/\sigma) + O(\sigma^{-3}) + O\left(\frac{1}{\sigma}e^{-\sigma A_j/[(1+\epsilon)m]}\right), \tag{12}$$

where $I(j) = |\mathcal{S}_1 \cap \mathcal{Z}(j)\setminus\{j\}|$, $J(j) = |\mathcal{S}_0 \cap \mathcal{Z}(j)\setminus\{j\}|$ and, for integers $I, J \geq 0$,

$$C_{I,J}^{(1)} = \int_{-\infty}^{\infty} h'(t)h(t)^I\{1 - h(t)\}^J dt, \quad C_{I,J}^{(2)} = \int_{-\infty}^{\infty} th'(t)h(t)^I\{1 - h(t)\}^J dt$$
$$C_{I,J}^{(3)} = \int_{-\infty}^{\infty} \{h'(t)\}^2 h(t)^I\{1 - h(t)\}^J dt.$$

(ii) If j is such that $\mathcal{N}(j) \nsubseteq \mathcal{S}_0$ or $\mathcal{P}(j) \nsubseteq \mathcal{S}_1$, then the derivative satisfies

$$\frac{\partial P(Y|\theta, \zeta)}{\partial \eta_j} = O\left(\frac{1}{\sigma}e^{-\sigma B_j/(1+\epsilon)}\right). \tag{13}$$

Moreover, if the outcome is neither increasing nor quasi-increasing, then (13) holds for all j.

Theorem 3 (Importance of quasi-increasing outcomes). *Quasi-increasing outcomes contribute terms of order $O(\sigma^{-1})$ or $O(\sigma^{-2})$ to the information matrix in (6). Outcomes that are not quasi-increasing contribute terms of order $O\left(\frac{1}{\sigma}e^{-\sigma \min_j B_j/(1+\epsilon)}\right)$ which are asymptotically negligible.*

Proofs for Theorems 1–3 are in Appendix 2. From Theorem 3, an asymptotic approximation to the information matrix in (6) need only include contributions from increasing and quasi-increasing outcomes. The importance of quasi-increasing outcomes seems difficult to capture using approximations, such as those in Sections 3.1–3.3, that only incorporate the first and second order moments of the joint distribution of the responses.

We combine the asymptotic approximations to $P(Y|\theta, \zeta)$ and $\partial P(Y|\theta, \zeta)/\partial \eta_j$, from Theorems 1 and 2 respectively, with (6) to provide an *asymptotic outcome-enumeration* approximation to the information matrix. The only additional requirement is that for the jth treatment in the ith support block ($i = 1, \ldots, b$; $j = 1, \ldots, m$), we must define a suitable partition of the indices $\{1, \ldots, m\}$ into sets $\mathcal{N}_i(j)$, $\mathcal{Z}_i(j)$, and $\mathcal{P}_i(j)$. This partition should be such that linear predictor $\eta_{ij'}$, relative to σ, is close to, less than, or greater than η_{ij} for $j' \in \mathcal{Z}_i(j)$, $j' \in \mathcal{N}_i(j)$ or $j' \in \mathcal{P}_i(j)$ respectively. We propose to form these partitions automatically using the heuristic algorithm given in Appendix 1. As presently implemented, the objective function corresponding to the asymptotic approximation is discontinuous; nonetheless, the resulting designs typically have high efficiencies relative to designs from the naïve outcome-enumeration approximation, competitive with those from the other methods. In certain circumstances, discussed in Section 5.2, this asymptotic approximation outperforms other methods.

When σ^2 is large, the recovery of inter-block information that occurs when using a mixed model for analysis is important for parameter estimation. For large σ^2, separation of outcomes (Albert & Anderson 1984) occurs within all blocks with high probability, in which case the parameters of the corresponding fixed block effects model are not estimable (see Propositions 1 and 2 in Appendix 2). Despite this, efficient parameter estimation is still possible under the mixed model (see Appendix 3).

4.2 Interpolated outcome-enumeration

In this section, we discuss a more direct numerical approximation for $M_\beta(\xi, \theta)$ under the logistic random intercept model. Note that for this model, the likelihood depends on the regression parameters β only through the vector η. Let $P_Y(\eta, \sigma^2) = P(Y|\zeta, \theta)$ and define

$$\mathcal{Q}(\eta, \sigma^2) = \sum_{Y \in \{0,1\}^m} \frac{1}{P_Y} \left(\frac{\partial P_Y}{\partial \eta} \right) \left(\frac{\partial P_Y}{\partial \eta} \right)^{\mathrm{T}}, \tag{14}$$

so that, by (6), $M_\beta(\zeta, \theta) = F^{\mathrm{T}} \mathcal{Q} F$.

An *interpolated outcome-enumeration* approximation to $M_\beta(\zeta, \theta)$ can be developed by surrogate modelling of the matrix-valued function $\mathcal{Q}$. The idea is to compute the values of the function $\mathcal{Q}$ at a collection of training points, and interpolate these data to predict the value of $\mathcal{Q}$ at new sites (η, σ^2). Interpolating $\mathcal{Q}$ as a function of η is particularly computationally efficient for finding Bayesian designs, as the same interpolator can be used for any value of β.

Surrogate modelling is widely applied in 'computer experiments' on expensive-to-evaluate computational models for complex phenomena (see Santner et al. 2003). We believe that its use for accelerating the computation of approximations necessary for the optimal design of physical experiments is new. For computer experiments, Gaussian process modelling (Kriging) is well-established as a surrogate; it can be used with training sets not arranged in a regular grid and can straightforwardly be applied to multidimensional problems. For block size $m = 2$, it is faster to use bilinear or bicubic interpolation and a regular grid.

5 Examples for binary response

5.1 Preliminaries for the examples

In Sections 5.2 and 5.3, D-optimal designs are found, compared and assessed for blocks of size $m = 4$ and a binary response logistic random intercept model with two variables and the following linear predictor

$$\nu(x; u, \beta) = \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} + u, \quad u \sim N(0, \sigma^2), \tag{15}$$

where $x = (x^{(1)}, x^{(2)})^{\mathrm{T}} \in [-1, 1]^2$. In Section 5.4, D-optimal designs are found for a logistic random intercept model with four factors and eight fixed parameters.

In Sections 5.2 and 5.4, we find locally D-optimal designs for various parameter scenarios by approximating the information matrix using adjusted generalized estimating equations and adjusted marginal quasi-likelihood. In Section 5.2 we also find locally optimal designs using unadjusted generalized estimating equations (assuming $\beta^* = \beta$) and, for large σ^2, asymptotic outcome-enumeration. In these sections we find it advantageous to specify parameter scenarios on the scale of the marginal effects, β_{att}, to facilitate performance comparisons across different values of σ^2. Intuitively, this setup mimics strong information being available for the marginal effects, and uncertainty in the strength of dependence. In Section 5.3, we find Bayesian D-optimal designs, with the prior information specified on the conditional parameters, as no comparisons are made across different values of σ^2; we set $\beta_0 \sim U[-0.5, 0.5]$, $\beta_1 \sim U[3, 5]$, $\beta_2 \sim U[0, 10]$, and $\sigma^2 = 5$. Thus, there is substantial uncertainty in the value of β_2, and moderate block-to-block variability. Here, we approximate the information matrix using the adjusted marginal quasi-likelihood, adjusted generalized estimating equations, and interpolated outcome-enumeration methods. For all of our examples, efficiencies of optimal designs found using the different approximations are calculated relative to D-optimal designs found using the naïve outcome-enumeration approximation.

For all approximations, we use a quasi-Newton method (the Broyden–Fletcher–Goldfarb–Shanno algorithm; Nocedal & Wright 1999, pp. 136–143) to obtain optimal, or near-optimal, designs numerically; that is optimal or highly efficient combinations of ζ_k, w_k and b. Multiple random starts of the algorithm are used to attempt to identify a global optimum of the objective function. Convergence is assessed via comparison of the optima obtained from the different starts, and was considered satisfactory for the examples presented here. We assess performance of the obtained designs using local efficiency, $\mathrm{eff}(\xi|\theta) = \{|M_\beta(\xi, \theta)|/\sup_{\xi'} |M_\beta(\xi', \theta)|\}^{1/p}$.

5.2 Example 1: Locally optimal designs

The purpose of this example is twofold. Firstly, we wish to illustrate the performance of the methods for different σ^2. Secondly, we demonstrate circumstances under which the resulting designs are robust to a reasonable range of values assumed for σ^2.

Table 1 gives the efficiencies under this regime of optimal designs from the different approximations relative to an optimal design found using the naïve outcome-enumeration approximation. It is clear that the unadjusted generalized estimating equation approach is by far the worst method, with efficiencies frequently less than 90%. In most cases, the remaining closed-form approximations are competitive with naïve outcome-enumeration. The performance of the adjusted generalized estimating equation approach depends critically on the choice of ρ which is treated here as a tuning parameter.

We observed two cases for which the adjusted marginal and adjusted generalized estimating equation methods performed poorly. For $\beta_{\mathrm{att}} = (1, 2, 3)^{\mathrm{T}}$ and $(2, 1, 3)^{\mathrm{T}}$, with $\sigma^2 = 50$, the design efficiencies from the former two methods were below 92%. These cases are unusual in that, for all $\sigma^2 > 1$, the two marginal approximations selected designs that replicate treatments within at least one of their blocks. This appears inefficient: the designs from both the naïve and asymptotic outcome-enumeration approximations do not feature within-block replication, and the latter design is at least 98% efficient. The only other case where this replication occurred in the marginal approximation designs for large σ^2 was $\beta_{\mathrm{att}} = (1, 2, 2)^{\mathrm{T}}$, where the efficiency was again relatively low. Our theoretical results (Section 4.1) suggest that marginal methods may poorly approximate the information matrix for designs featuring within-block replication when σ^2 is large. Thus we would recommend some caution when σ^2 is large and use of the marginal approximations yields designs featuring within-block replication of treatments. For such designs, the error from these approximations may be large. Additionally, the small u Taylor approximations underlying the covariance approximation in the adjusted marginal quasi-likelihood method cannot be expected to be accurate when σ^2 is large and large random effects are anticipated.

Table 2 gives the average total processor time for each method, as recorded in a high performance parallel computing environment with twelve 2·4GHz cores per node. The times given are per parameter vector for 100 random starts of the optimization algorithm. Naïve outcome-enumeration is the most expensive method followed by asymptotic outcome-enumeration, adjusted generalized estimating equations and adjusted marginal quasi-likelihood. The computational expense of the adjusted generalized estimating equations method depends on the structure of the problem. Here, there are many parameter scenarios with the same values of β_{att} which allows re-use of adjusted generalized estimating equation designs for a given β_{att} for various σ^2. If re-use were not possible, then the time per design would be higher: an indicative figure is given in parentheses. The time to obtain a design for given ρ is comparable

		σ^2					
$\beta_{\mathrm{att}}^{\mathrm{T}}$	Design	1	2	5	10	20	50
(0,1,1)	Unadj. gen.	96·3–100·0	94·5–100·1	82·9–99·6	78·8–94·3	71·8–87·1	59·9–76·6
	Adj. marg.	100	100	100	100	100	100
	Adj. gen.	99·7–100·0	99·7–100·1	99·2–100·0	98·8–100·0	98·5–100·0	98·1–100·0
	Asymp. enum.					100·0	94·8
(0,3,2)	Unadj. gen.	86·2–97·3	84·5–93·2	79·1–85·3	74·7–79·0	70·9–73·4	63·3–67·7
	Adj. marg.	99·9	99·9	100·0	99·9	99·4	95·2
	Adj. gen.	85·3–99·8	85·6–99·6	86·3–99·5	87·2–99·7	87·7–100·0	83·9–98·5
	Asymp. enum.					96·4	97·4
(0,5,10)	Unadj. gen.	82·3–96·1	79·8–91·9	70·1–84·4	65·2–78·4	64·7–72·8	52·2–67·0
$\times(1+5c^2)^{-\frac{1}{2}}$	Adj. marg.	99·9	99·9	100·0	99·8	99·7	99·8
	Adj. gen.	83·9–99·1	84·1–98·7	84·8–98·9	85·6–99·4	86·0–99·5	83·9–99·1
	Asymp. enum.					94·8	96·1
(1,2,3)†	Unadj. gen.	84·3–96·8	83·2–94·5	75·2–88·5	70·6–78·7	65·7–78·9	58·9–73·9
	Adj. marg.	100·4	99·1	96·6	92·1	84·8	73·5(*)
	Adj. gen.	84·1–99·5	85·0–99·0	86·6–99·2	87·6–98·9	86·7–97·2	77·6–91·7(*)
	Asymp. enum.					93·5	98·3
(1,4,4)	Unadj. gen.	81·7–96·9	80·7–94·1	76·0–86·3	70·8–79·8	65·0–73·1	58·3–64·6
	Adj. marg.	100·0	100·0	99·9	99·4	98·2	97·2
	Adj. gen.	80·3–99·4	81·0–99·1	81·9–99·3	82·5–99·7	82·5–98·9	82·5–97·4
	Asymp. enum.					97·1	98·2
(1,3,3)	Unadj. gen.	82·2–97·1	80·6–93·5	76·4–86·4	71·0–79·7	65·0–72·7	57·1–63·1
	Adj. marg.	99·7	99·5	100·0	99·3	97·9	95·1
	Adj. gen.	79·6–99·3	80·5–98·6	83·1–99·4	84·9–99·7	85·9–98·8	84·6–95·6
	Asymp. enum.					97·1	98·2
(1,2,2)	Unadj. gen.	82·1–97·9	81·9–93·4	78·6–88·2	73·4–78·6	67·3–69·5	58·4–64·3
	Adj. marg.	100·6	100·3	100·2	98·5	96·1	92·6
	Adj. gen.	83·3–100·1	84·6– 99·3	87·4– 99·8	89·3– 99·6	90·1– 98·0	88·5– 95·4
	Asymp. enum.					95·1	97·9
(2,1,3)†	Unadj. gen.	84·3–96·8	83·2–94·5	78·5–89·2	73·4–85·9	62·8–76·7	56·4–64·1
	Adj. marg.	99·9	99·1	96·6	92·1	84·8	78·0(*)
	Adj. gen.	83·6-99·5	84·5-99·0	86·1-99·1	86·8-98·9	85·8-97·0	77·4-90·9(*)
	Asymp. enum.					95·0	98·0

Table 1: Example 1: computed efficiencies of locally D-optimal designs from different methods. Unadj. gen. – Unadjusted generalized estimating equations; Adj. marg. – adjusted marginal quasi-likelihood; Adj. gen. – adjusted generalized estimating equations; Asymp. enum. – asymptotic outcome-enumeration. Reported efficiencies for generalized estimating equation methods are for $\rho = 0\cdot1$, $0\cdot15$, $0\cdot2$, …, $0\cdot7$. Symbols † and (*) indicate parameter values for which the adjusted marginal modelling approximations give particularly inefficient designs.

Method	Time per parameter vector (processor-seconds)
Naïve outcome-enumeration ($\sigma^2 = 50$)	3×10^5
Naïve outcome-enumeration ($\sigma^2 = 1$)	5×10^4
Asymptotic outcome-enumeration	7×10^3
Adjusted generalized estimating equations*	1×10^3 (8×10^3)
Adjusted marginal quasi-likelihood	4×10^2

Table 2: Example 1: computational expense for locally D-optimal designs. *Figure in brackets is indicative of time when design re-use is impossible.
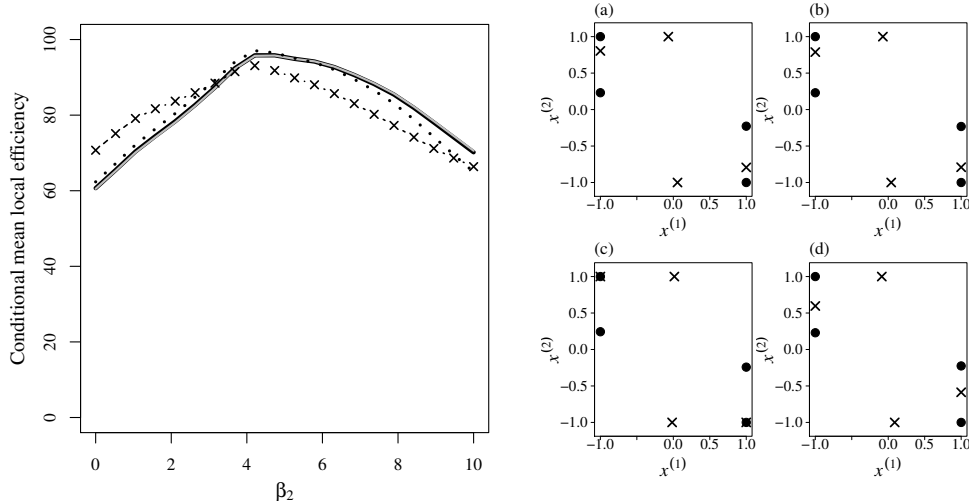
Figure 1: Example 2: conditional mean efficiencies $E[\text{eff}(\xi;\theta)|\beta_2]$ (left panel) and support blocks (right panel) of Bayesian D-optimal designs from maximum likelihood via naïve outcome-enumeration [black line, (a)], maximum likelihood via interpolated outcome-enumeration [solid grey line, (b)], adjusted marginal quasi-likelihood [dotted black line, (c)], and adjusted generalized estimating equations [dashed and crossed black line, (d)]. Treatments with the same plotting character are in the same block.

with that from adjusted marginal quasi-likelihood.

For moderate dependence ($\sigma^2 \leq 10$) choosing a single value of $\sigma^2 = 5$ appears to be very robust; for all β_{att} considered, the naïve outcome-enumeration design with $\sigma^2 = 5$ has a calculated efficiency of at least 99·1% for $\sigma^2 = 1, 2, 10$. Assuming a single value of $\sigma^2 = 1$ is less robust, though still reasonable, the worst case is when $\beta_{\text{att}} = (1, 2, 2)^{\text{T}}$ and the true $\sigma^2 = 10$; the efficiency of the resulting design is 97·1%. However, if the dependence is actually strong then the above designs may perform comparatively poorly; when $\beta_{\text{att}} = (0, 3, 2)^{\text{T}}$, $\sigma^2 = 50$, the design obtained assuming $\sigma^2 = 5$ has a calculated D-efficiency of 93·6%. This robustness of an optimal design to a wide range of assumed values of σ^2 is a consequence of specifying the parameters on the marginal scale.

5.3   Example 2: Bayesian optimal designs

We choose the design ξ to maximize $\psi(\xi)$ from Section 2.3 and approximate the integral in the objective function by averaging over a Latin hypercube sample of 50 values of β from $[-0.5, 0.5] \times [3, 5] \times [0, 10]$. As the value of σ^2 is assumed known, for the interpolated outcome-enumeration approximation we only need build a surrogate model of $\mathcal{Q}$ as a function of η.

Bayesian D-optimal designs were computed for each of the different approximations using 1000 random starts; the support blocks of the designs are shown in Fig. 5.3 with corresponding weights given in Table 3. A single value $\rho = 0.6$, corresponding to fairly strong correlation, was used in the adjusted generalized estimating equations approximation. For each method, from the 1000 designs generated the best was selected with respect to naïve outcome-enumeration.

All of the designs contain multiple support blocks due to the degree of uncertainty in the parameters and the small block size. Locally D-optimal designs were also found for each of the 50 sampled parameter vectors under the naïve outcome-enumeration approximation, and the local efficiency of each Bayesian design was calculated relative to each of these 50 designs. Then, Gaussian process emulators were constructed for the efficiency profile of each Bayesian design. Figure 5.3 shows the dependence of the efficiency on β_2, via approximations of $E[\text{eff}(\xi;\theta)|\beta_2]$ obtained from the efficiency profile emulators. The performance of all of the Bayesian designs varied little according to the value of β_0 or β_1, with $E[\text{eff}(\xi;\theta)|\beta_0]$ and $E[\text{eff}(\xi;\theta)|\beta_1]$ changing by fewer than 4 percentage points over the ranges of β_0 and β_1 respectively. The conditional mean efficiency of the design from the adjusted generalized estimated equations approach is clearly quite different, as a function of β_2, from the local efficiencies from the other methods. The designs from all of the approximations appear similar to the naïve outcome-enumeration

design (compare Figures 5.3(a)–(d)).

To train the interpolated outcome-enumeration approximation of $\mathcal{Q}$, a random Latin hypercube sample of 10,000 η vectors was drawn from $[-20, 20]^4$, and the matrix $\mathcal{Q}$ evaluated for each vector. The second-order, compactly-supported Wendland covariance function was used, with range parameter chosen manually as 15 to make the predictions appear reasonably smooth and accurate. Independent Gaussian process models were fitted to the m^2 entries of $\mathcal{Q}$. The use of a compactly-supported covariance function is advantageous here due to the large number of training points; it enables inversion of the covariance matrix in a reasonable time, and permits relatively fast predictions from the fitted model. For finding Bayesian designs, the interpolation method required around 3·2 times less computational effort than naïve outcome-enumeration for this example (Table 3). If more quadrature points were used to approximate the prior distribution, or if an adequate emulator could be found using fewer training points, then the advantage of using interpolation to approximate the objective function would be greater (for 200 quadrature points, with the same training set, objective function evaluation using interpolation is approximately 6 times faster than naïve outcome-enumeration). The advantage will also be more pronounced for larger σ^2. The closed-form approximations (using a single ρ) are approximately two orders of magnitude faster than naïve outcome-enumeration.

| | Block weights | | | |
Design method	•	×	Bayes efficiency	Time (processor-seconds)
Likelihood, naïve outcome-enumeration	0·744	0·256	100·00	$1·65\times10^7$
Likelihood, interpolated outcome-enumeration	0·749	0·251	99·96	$5·19\times10^6$
Adjusted marginal quasi-likelihood	0·748	0·252	99·79	$1·80\times10^5$
Adjusted estimating equations	0·466	0·534	97·94	$2·20\times10^5$

Table 3: Example 2: details of Bayesian designs . Above, • and × correspond to symbols in Fig. 5.3(a)–(d). The Bayes efficiency of ξ is $\exp[\{\psi(\xi) - \sup_{\xi'} \psi(\xi')\}/p]$.

5.4 Example 3: Locally optimal designs, four factors

We investigated locally optimal designs with $x = (x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)})^{\mathrm{T}} \in [-1, 1]^4$, and

$$
\begin{aligned}
\nu(x; u, \beta) =& \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \beta_3 x^{(3)} + \beta_4 x^{(4)} \\
& + \beta_{12} x^{(1)} x^{(2)} + \beta_{13} x^{(1)} x^{(3)} + \beta_{14} x^{(1)} x^{(4)} + u, \quad u \sim N(0, \sigma^2),
\end{aligned}
$$

with $\beta_{\mathrm{att}} = (2, 3, 0, 3, 0, 0, -2, 0)^{\mathrm{T}}$, $(1, 2, 1, -3, -1, \frac{1}{4}, -\frac{1}{2}, 3)^{\mathrm{T}}$, $(0, 1, 1, 1, 1, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})^{\mathrm{T}}$, and $\sigma^2 = 1, 2, 5$. Designs were found using the naïve outcome-enumeration, adjusted marginal quasi-likelihood and adjusted generalized estimating equation ($\rho = 0·3, 0·5, 0·6$) approximations with 100, 1000 and 1000 random starts respectively. In all cases, the marginal approximations required less computational effort despite the more thorough search, yielding designs with at least 99.5% efficiency relative to the design from the naïve outcome-enumeration approximation.

6 Poisson response

6.1 Approach

In this section we demonstrate the use of the marginal quasi-likelihood approximation to find D-optimal designs for a Poisson model with random intercept. We compare the designs to those of Niaparast (2009), who investigated design for this model using a direct quasi-likelihood approximation to the information matrix, and also to the designs from the analytical results of Russell et al. (2009) for the Poisson model with no random effects. The conditional distribution of the response is assumed to be Poisson, with link function $g(\mu) = \log(\mu)$. In the random intercept model, $u \sim N(0, \sigma^2)$ is a scalar, and $\nu(x; u, \beta) = f^{\mathrm{T}}(x)\beta + u$.

Quasi-likelihood estimation requires a parametric specification of only the marginal mean and variance of the response, and not a full probability model. Niaparast (2009) obtained a covariance matrix for the resulting parameter estimators using the actual marginal mean and variance for the Poisson random
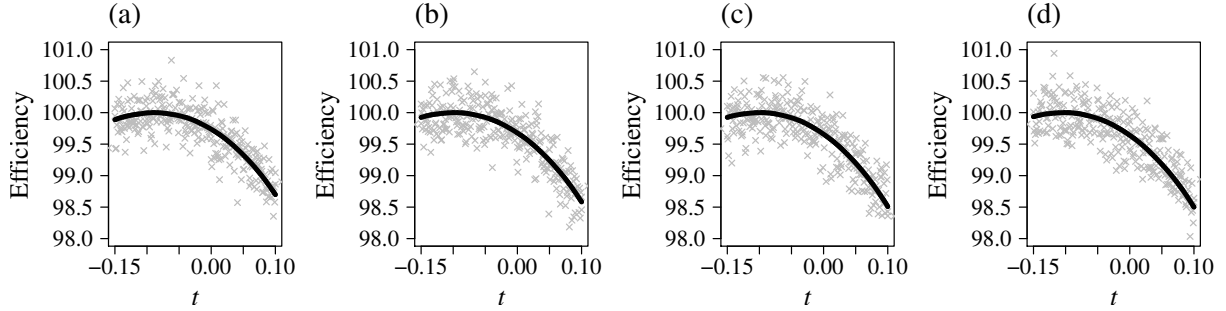
Figure 2: Approximate efficiency of the design $\zeta(t)$ obtained using Monte Carlo approximation and nonparametric smoothing, for the following values of σ^2: (a) 0·01, (b) 0·025, (c) 0·05, (d) 0·1.

intercept model which are analytically tractable. We shall refer to this as the 'direct' approach. In general, there are issues with the use of quasi-likelihood for dependent data (McCullagh & Nelder 1989, Ch.9); however the above approach could be viewed as an application of generalized estimating equations (Liang & Zeger 1986) with a working correlation structure calculated from the full probability model.

6.2 Comparison of designs, $m = 3$

Locally D-optimal designs for the Poisson random intercept model were computed by numerically optimizing the determinant of the information matrix under the marginal quasi-likelihood and direct, quasi-likelihood, approximations. The linear predictor structure (15) was assumed, with conditional parameter values $(\beta_0, \beta_1, \beta_2) = (3, 1, 2)$, together with several values for σ^2.

For $\sigma^2 = 0$, the designs found numerically coincided with those anticipated by the theoretical results of Russell et al. (2009) for models with no random effects. For $\sigma^2 = 0·01, 0·025, 0·05, 0·1$, each of the designs contains a single support block ($b = 1$, $w_1 = 1$) of the form $\zeta(t) = ((1, 1)^{\mathrm{T}}, (-1, 1)^{\mathrm{T}}, (1, t)^{\mathrm{T}})$, with t = -0·083, -0·091, -0·095, -0·096 respectively. The designs from the two methods agree to three decimal places.

We assess the efficiency, for maximum likelihood estimation, of designs resulting from the choice of $t \in [-1, 1]$ by using Monte Carlo integration to approximate $M_\beta(\zeta(t); \theta)$ (Section 2.2) and nonparametric smoothing to obtain a surrogate, $\tilde{\Psi}(t)$, for $\Psi(t) = |M_\beta(\zeta(t); \theta)|$ (see also Müller & Parmigiani 1995). Let $t^\star = \max_{t' \in [-1,1]} \tilde{\Psi}(t')$. Figure 2 shows the approximate efficiency in the neighbourhood of the optimal t, obtained from $\tilde{\mathrm{eff}}(t) = \{\tilde{\Psi}(t)/\tilde{\Psi}(t^\star)\}^{1/p}$, together with estimates of $\{\Psi(t)/\tilde{\Psi}(t^\star)\}^{1/p}$ each using 10^5 Monte Carlo samples. The total processor time for the Monte Carlo computations was approximately $1·5 \times 10^5$s, using a sixteen-core 2·6 GHz node. The results indicate that, for all values of σ^2 considered here, both the marginal and direct quasi-likelihood designs have an efficiency around 100%, and also any choice of t in [-0·15, 0] will be very highly efficient.

The direct, quasi-likelihood, approach for the Poisson response is similar to the adjusted marginal or adjusted generalized estimating equations methods for a binary response, in the sense that it accounts for the form of the marginal mean. Theoretically, it has the advantage of not relying on Taylor series approximations. Unlike for binary data, where an unadjusted marginal method is poor, for a Poisson response the unadjusted method has virtually identical performance to the direct method. This is perhaps to be expected if we consider that the normal approximations to the response distribution used in the marginal approximation are much more accurate for Poisson than binary responses. There is essentially no computational advantage to the Taylor-series based approximation and so we would recommend the direct approach as a default first choice.

Note the values of σ^2 used here are much smaller than those used for binary response models in Section 5; for a Poisson response, σ^2 is chosen to give a plausible range of marginal overdispersion over $\mathcal{X}$ (for example, approximately 1·10–45·6 when $\sigma^2 = 0·1$), and plausible correlation between responses from units in the same block receiving the same treatment.

12

7 Discussion

For the logistic random intercept model, use of a correction for the marginal attenuation of the parameters yields much improved designs; in our examples, designs using this idea often performed on a par with those from naïve outcome-enumeration. Further investigations, including simulations to assess small sample properties, are available in the first author's Ph.D. thesis.

Tekle et al. (2008) employed an information matrix approximation derived from penalized quasi-likelihood (Breslow & Clayton 1993). Their approach requires predictions of the random effects, which they approximated at the design stage using Monte Carlo simulation. The resulting approximation is computationally intensive and is not suitable for routine use on more complex problems. Hence, we chose not to pursue this methodology here.

Avenues for future research include developing the necessary methodology to extend the adjusted closed-form approximations to find designs for models with more complex random effects, and extension of the asymptotic results in Section 4.1 to other link functions for binary response.

Acknowledgements

Appendix 1: An algorithm for determining membership of sets $\mathcal{N}(j)$, $\mathcal{Z}(j)$, $\mathcal{P}(j)$

To obtain an asymptotic approximation that performs reasonably for a broad choice of designs, a decision is required on which η_j should be considered 'close'; that is, for given ζ and j, which indices should we treat as belonging to $\mathcal{Z}(j)$ in order to apply Theorems 1–3? Below we give the algorithm used in Example 1. The algorithm approximates exponentially decaying error terms as zero.

For calculation of the derivatives, the algorithm iteratively augments $\mathcal{Z}(j)$ with the index, l, of the next closest predictor to η_j if two conditions are satisfied. Condition (A) concerns the coefficient of $\phi(-\eta_j/\sigma)/\sigma$ in the expression of Theorem 2, which is an approximation to an integral of the form $\int_{-\infty}^{\infty} h'(t) f_2(t, \sigma^2) dt$, see equation (A3) in Appendix 2. The value of this integral decreases as the set $\mathcal{Z}(j)$ is augmented. Condition (B), concerning the same coefficient, is a heuristic that prevents the application of a Taylor approximation when Δ_{lj} is too large, see (A4) in Appendix 2.

For the probability calculation, we use the expression from part (i) of Theorem 1 unless $\max_{j \in \mathcal{S}_0}\{\eta_j\}$ and $\min_{j \in \mathcal{S}_1}\{\eta_j\}$ are close (less than 1 apart), in which case we take $j' = \arg\max_{\{j \in \mathcal{S}_0\}}\{\eta_j\}$, $l' = \arg\min_{\{j \in \mathcal{S}_1\}}\{\eta_j\}$, $\mathcal{Z}(j') = \{j', l'\}$ and use the expression in part (ii) of Theorem 1. The cutoff distance of $\gamma = 1$ is chosen because at this point the probabilities in parts (i) and (ii) should be similar, since $\Phi(-\eta_j/\sigma) - \Phi(-\eta_l/\sigma) \approx \frac{\eta_l - \eta_j}{\sigma}\phi(-\eta_j/\sigma) = \frac{1}{\sigma}\phi(-\eta_j/\sigma)$.

Algorithm 1. *For each possible outcome Y, approximate its contribution, $\frac{1}{P(Y)}\left\{\frac{\partial P(Y)}{\partial \eta}\right\}\left\{\frac{\partial P(Y)}{\partial \eta}\right\}^{\mathrm{T}}$, to the information matrix in (6) using Theorems 1–3 to approximate $P(Y)$ and $\partial P(Y)/\partial\eta$, and add it to the total.*
To compute $P(Y)$:

  Compute $\lambda_0 = \max_{j \in \mathcal{S}_0}\{\eta_j\}$ and $\lambda_1 = \min_{j \in \mathcal{S}_1}\{\eta_j\}$
  If $\lambda_1 \geq \lambda_0 + \gamma$:
    Set $P(Y) \leftarrow \Phi(\lambda_1/\sigma) - \Phi(\lambda_0/\sigma)$ [using Theorem 1(i)]
  If $|\lambda_1 - \lambda_0| \leq \gamma$:
    Set $P(Y) \leftarrow \frac{\phi(\lambda_1/\sigma)}{\sigma}$ [using Theorem 1(ii)]
  If $\lambda_1 \leq \lambda_0 - \gamma$, set $1/P(Y) \leftarrow 0$, and do not compute $\partial P(Y)/\partial\eta_j$
    i.e. do not include a contribution from this outcome in the information matrix approximation

To compute $\partial P(Y)/\partial\eta_j$:

  Declare $\mathcal{Z}(j) = \{j\}$

Set $C_4 = 1$

Propose augmenting $\mathcal{Z}(j)$ to $\mathcal{Z}'(j) = \{j, \arg\min_{l \neq j} |\eta_j - \eta_l|\}$

Iterate until STOP. Given current proposal $\mathcal{Z}'(j)$:

> *Calculate $I(j), J(j)$ for $\mathcal{Z}'(j)$, refer to as I', J' respectively*
>
> *Set $C_4' \leftarrow C_{I',J'}^{(1)} + C_{I'-1,J'}^{(3)} \sum_{l \in \mathcal{S}_1 \cap \mathcal{Z}'(j) \setminus \{j\}} \Delta_{lj} - C_{I'-1,J'-1}^{(3)} \sum_{l \in \mathcal{S}_0 \cap \mathcal{Z}'(j) \setminus \{j\}} \Delta_{lj}$*
>
> *If (A) $0 \leq C_4' \leq C_4$ and (B) $|C_4' - C_{I',J'}^{(1)}| \leq |C_4 - C_{I',J'}^{(1)}|$, accept proposal*
>> *Update $C_4 \leftarrow C_4'$, $\mathcal{Z}(j) \leftarrow \mathcal{Z}'(j)$*
>
> *If did not accept proposal in previous step, then STOP*
>
> *Otherwise make new proposal, $\mathcal{Z}'(j) \leftarrow \mathcal{Z}(j) \cup \arg\min_{l \notin \mathcal{Z}(j)}\{|\eta_l - \eta_j|\}$*

Set $\mathcal{N}(j) \leftarrow \{l : \eta_l < \eta_{l'}, \text{ for all } l' \in \mathcal{Z}(j)\}$

Set $\mathcal{P}(j) \leftarrow \{l : \eta_l > \eta_{l'}, \text{ for all } l' \in \mathcal{Z}(j)\}$

If $\{\mathcal{S}_1 \cap \mathcal{N}(j)\} \cup \{\mathcal{S}_0 \cap \mathcal{P}(j)\} = \emptyset$:

> *Deem the outcome as quasi-increasing*
>
> *Set $\frac{\partial P(Y)}{\partial \eta_j} \leftarrow (2y_j - 1) \max\left\{0, \frac{1}{\sigma}\phi\left(\frac{-\eta_j}{\sigma}\right) C_4 + \frac{1}{\sigma^2}\phi'\left(\frac{-\eta_j}{\sigma}\right) C_{I(j),J(j)}^{(2)}\right\}$ [using Theorem 2(i)]*

Else set $\frac{\partial P(Y)}{\partial \eta_j} \leftarrow 0$ [using Theorem 2(ii)]

Appendix 2: Proofs and further asymptotic results

Recall that $\mathcal{Z}(j) = \{l : \eta_l - \eta_j \to 0\}$, $\mathcal{N}(j) = \{l : \eta_l - \eta_j \to -\infty\}$, $\mathcal{P}(j) = \{l : \eta_l - \eta_j \to \infty\}$, and $\mathcal{S}_0 = \{j : y_j = 0\}$, $\mathcal{S}_1 = \{j : y_j = 1\}$. For the asymptotic results, we require some assumptions repeated here for clarity.

Assumption 4. $\beta_{att} = \beta/\sqrt{1 + c^2\sigma^2}$ is fixed as $\sigma^2 \to \infty$.

Assumption 5. For all $j = 1, \ldots, m$, either $\eta_j^* = f^{\mathrm{T}}(x_j)\beta_{att}$ is fixed or there exists $l \neq j$ with η_l^* fixed and $\eta_l^* - \eta_j^* = o(\sigma^{-1})$.

Assumption 6. There exists $A_j > 0$ such that $|\eta_l - \eta_j| > \sigma A_j$ for $l \in \{\mathcal{S}_0 \cap \mathcal{N}(j)\} \cup \{\mathcal{S}_1 \cap \mathcal{P}(j)\}$, and $B_j > 0$ such that $|\eta_l - \eta_j| > \sigma B_j$ for all $l \in \{\mathcal{S}_1 \cap \mathcal{N}(j)\} \cup \{\mathcal{S}_0 \cap \mathcal{P}(j)\}$.

Define

$$f_{1,j}(t, \sigma^2) = \prod_{l \in \mathcal{S}_1 \cap \mathcal{N}(j)} h(\eta_l - \eta_j + t) \prod_{l \in \mathcal{S}_0 \cap \mathcal{P}(j)} \{1 - h(\eta_l - \eta_j + t)\}$$

$$f_{2,j}(t, \sigma^2) = \prod_{l \in \mathcal{S}_1 \cap \mathcal{Z}(j) \setminus \{j\}} h(\eta_l - \eta_j + t) \prod_{l \in \mathcal{S}_0 \cap \mathcal{Z}(j) \setminus \{j\}} \{1 - h(\eta_l - \eta_j + t)\}$$

$$f_{3,j}(t, \sigma^2) = \prod_{l \in \mathcal{S}_1 \cap \mathcal{P}(j)} h(\eta_l - \eta_j + t) \prod_{l \in \mathcal{S}_0 \cap \mathcal{N}(j)} \{1 - h(\eta_l - \eta_j + t)\}$$

We will mostly suppress the dependence of these functions on j and write f_1, f_2, f_3 where the context is clear. Fix $t \in \mathbb{R}$. If $\{\mathcal{S}_1 \cap \mathcal{N}(j)\} \cup \{\mathcal{S}_0 \cap \mathcal{P}(j)\} \neq \emptyset$, then $f_1(t, \sigma^2) \to 0$ as $\sigma^2 \to \infty$, otherwise $f_1(t, \sigma^2) = 1$. We always have $f_3(t, \sigma^2) \to 1$. We make use of the following lemma.

Lemma 1. *Suppose f_3 is as defined above, and $f_4(t, \sigma^2)$ is measurable as a function of t for all fixed σ^2, with $0 \leq f_4(t, \sigma^2) \leq K$ for all t, σ^2, for some $K > 0$. Then:*

(i) For any $\epsilon > 0$, as $\sigma^2 \to \infty$,

$$\int_{-\infty}^{\infty} h'(t) f_3(t, \sigma^2) f_4(t, \sigma^2)dt = \int_{-\infty}^{\infty} h'(t) f_4(t, \sigma^2)dt + O(e^{-\sigma A_j/[(1+\epsilon)m]}),$$

i.e. replacing f_3 by 1 in the integrand incurs only an exponentially decaying error.

(ii) Suppose that Δ_1, Δ_2 vary with σ^2, but $|\Delta_1|, |\Delta_2| \leq \Delta_{\max}$. Then, as $\sigma^2 \to \infty$, for any $\epsilon > 0$,

$$\int_{-\infty}^{\infty} h(t + \Delta_1)\{1 - h(t + \Delta_2)\} f_3(t, \sigma^2) f_4(t, \sigma^2)dt$$

$$= \int_{-\infty}^{\infty} h(t + \Delta_1)\{1 - h(t + \Delta_2)\} f_4(t, \sigma^2)dt + O(e^{-\sigma A_j/[(1+\epsilon)m]}),$$

i.e. the integrator, $h'(t)$, in (i) can be replaced by $h(t + \Delta_1)\{1 - h(t + \Delta_2)\}$.

The key idea in the proof of Lemma 1 is to approximate the logistic function by a step function. Observe that if h is the logistic function and $S(t) = \mathbb{I}(t > 0)$, then there is $L > 0$ such that $|h(t) - S(t)| \leq Le^{-|t|}$. Moreover, we can reduce the rate constant for the exponential and still have an upper bound. Thus, given $\epsilon > 0$, $|h(t) - S(t)| \leq Le^{-|t|/(1+\epsilon)}$.

As a prelude to the proof of Lemma 1, we demonstrate exponential convergence of a relatively simple integral to zero. The full proof is more intricate, but does not involve many more ideas. Observe

$$
\begin{aligned}
\left| \int_{-\infty}^{\infty} [h(t+\sigma) - S(t+\sigma)] h'(t) dt \right| &\leq \int_{t+\sigma>0} Le^{-\sigma/(1+\epsilon) - t/(1+\epsilon)} h'(t) dt + \int_{t+\sigma<0} h'(t) dt \\
&\leq Le^{-\sigma/(1+\epsilon)} \int_{-\infty}^{\infty} e^{-t/(1+\epsilon)} h'(t) dt + h(-\sigma) \\
&= O(e^{-\sigma/(1+\epsilon)}).
\end{aligned}
$$

Key to the conclusion is the observation that the integral in the second line is finite. This is true since in the upper and lower tails the integrand is bounded, respectively, by $\lambda e^{-|t|\{1+1/(1+\epsilon)\}}$ and $\lambda e^{-|t|\{1-1/(1+\epsilon)\}}$, where $\lambda > 1$. The integral is not finite if $\epsilon = 0$.

of Lemma 1. Part (i): Observe that

$$
f_3(t, \sigma^2) = \prod_{l \in \mathcal{S}_0 \cap \mathcal{N}(j)} h(-(\eta_l - \eta_j + t)) \prod_{l \in \mathcal{S}_1 \cap \mathcal{P}(j)} h(\eta_l - \eta_j + t).
$$

Assume $A_j \sigma + t > 0$. Then, for $l \in \mathcal{S}_1 \cap \mathcal{P}(j)$, there is a constant $L_m > 0$ such that

$$
\begin{aligned}
|h(\eta_l - \eta_j + t) - 1| = |h(\eta_l - \eta_j + t) - S(\eta_l - \eta_j + t)| &\leq L_m e^{-|\eta_l - \eta_j + t|/((1+\epsilon)m)} \\
&\leq L_m e^{-\sigma A_j/((1+\epsilon)m) - t/((1+\epsilon)m)} \leq L_m e^{-\sigma A_j/((1+\epsilon)m) + |t|/((1+\epsilon)m)}.
\end{aligned}
$$

By a similar argument, L_m can also be chosen such that, in addition, for $l \in \mathcal{S}_0 \cap \mathcal{N}(j)$ and $t < A_j \sigma$,

$$
|h(-(\eta_l - \eta_j + t)) - 1| \leq L_m e^{-A_j \sigma/((1+\epsilon)m) + |t|/((1+\epsilon)m)}.
$$

Thus, for $-A_j \sigma < t < A_j \sigma$,

$$
\prod_{l \in \{\mathcal{S}_0 \cap \mathcal{N}(j)\} \cup \{\mathcal{S}_1 \cap \mathcal{P}(j)\}} \left\{ 1 - L_m e^{-\sigma A_j/((1+\epsilon)m) + |t|/((1+\epsilon)m)} \right\} \leq f_3(t, \sigma^2) \leq 1.
$$

Let $\kappa = |\{\mathcal{S}_0 \cap \mathcal{N}(j)\} \cup \{\mathcal{S}_1 \cap \mathcal{P}(j)\}|$, noting $\kappa \leq m$. Binomial expansion of the product yields a conservative bound,

$$
\begin{aligned}
\left| \int_{-A_j \sigma}^{A_j \sigma} h'(t) \{1 - f_3(t, \sigma^2)\} dt \right| &\leq \sum_{l=1}^{\kappa} \binom{\kappa}{l} L_m^l e^{-l\sigma A_j/((1+\epsilon)m)} \int_{-\infty}^{\infty} e^{l|t|/((1+\epsilon)m)} h'(t) dt \\
&= O(e^{-\sigma A_j/((1+\epsilon)m)}),
\end{aligned}
$$

as the integral on the right hand side is finite for $l \leq m$. Moreover,

$$
\left| \int_{A_j \sigma}^{\infty} h'(t) \{1 - f_3(t, \sigma^2)\} dt \right| \leq 1 - h(A_j \sigma) = O(e^{-A_j \sigma}),
$$

and similarly

$$
\left| \int_{-\infty}^{-A_j \sigma} h'(t) \{1 - f_3(t, \sigma^2)\} dt \right| \leq h(-A_j \sigma) = O(e^{-A_j \sigma}).
$$

Overall,

$$
\left| \int_{-\infty}^{\infty} h'(t) \{1 - f_3(t, \sigma^2)\} dt \right| = O(e^{-\sigma A_j/((1+\epsilon)m)}).
$$

When combined with the assumption $0 \leq f_4(t, \sigma^2) \leq K$, this is adequate to prove the lemma.

15

Part (ii): First note that there exists $K' > 0$ such that, for all Δ_1, Δ_2 with $|\Delta_1|, |\Delta_2| \le \Delta_{\max}$,

$$h(t + \Delta_1)\{1 - h(t + \Delta_2)\} \le K'h'(t). \tag{16}$$

Now consider the case $f_4 = 1$, for which we have

$$\int_{-\infty}^{\infty} h(t + \Delta_1)\{1 - h(t + \Delta_2)\}\{1 - f_3(t, \sigma^2)\}dt$$

$$\le K' \int_{-\infty}^{\infty} h'(t)\{1 - f_3(t, \sigma^2)\}dt = O(e^{-\sigma A_j/((1+\epsilon)m)}),$$

as established in part (i). The result for general f_4 holds via a similar argument to part (i).

It can be seen that a conservative choice in (16) above is $K' = 4 \exp \Delta_{\max}$. To show this, note

$$R := \frac{h(t + \Delta_1)\{1 - h(t + \Delta_2)\}}{h'(t)} = \frac{e^{\Delta_1}(1 + e^t)^2}{(1 + e^{t+\Delta_1})(1 + e^{t+\Delta_2})} = \frac{e^{\Delta_1}(e^{-t} + 1)^2}{(e^{-t} + e^{\Delta_1})(e^{-t} + e^{\Delta_2})}.$$

For $t \ge 0$, use the final expression above to see that $R \le e^{\Delta_1} \times 4/e^{\Delta_1 + \Delta_2} = 4e^{-\Delta_2} \le 4e^{\Delta_{\max}}$. For $t < 0$, considering the penultimate expression above we see $R \le e^{\Delta_1} \times 4/1 \le 4e^{\Delta_{\max}}$. $\square$

of Theorem 2 (Derivatives). Part (i): The derivative is given by

$$\frac{\partial P(Y)}{\partial \eta_j} = (2y_j - 1) \int_{-\infty}^{\infty} h'(\eta_j + \sigma u) \prod_{l \in \mathcal{S}_1 \setminus \{j\}} h(\eta_l + \sigma u) \prod_{l \in \mathcal{S}_0 \setminus \{j\}} \{1 - h(\eta_l + \sigma u)\}\phi(u)du$$

$$= \frac{(2y_j - 1)}{\sigma} \int_{-\infty}^{\infty} h'(t) f_1(t, \sigma^2) f_2(t, \sigma^2) f_3(t, \sigma^2) \phi\left(\frac{t}{\sigma} - \frac{\eta_j}{\sigma}\right) dt, \tag{17}$$

since, from Assumption 5, $\mathcal{N}(j) \cup \mathcal{Z}(j) \cup \mathcal{P}(j) = \{1, \ldots, m\}$. If $\mathcal{S}_1 \cap \mathcal{N}(j) = \mathcal{S}_0 \cap \mathcal{P}(j) = \emptyset$, then $f_1 = 1$ and, from Lemma 1(i), (17) is equal to

$$\frac{(2y_j - 1)}{\sigma} \int_{-\infty}^{\infty} h'(t) f_2(t, \sigma^2) \phi\left(\frac{t}{\sigma} - \frac{\eta_j}{\sigma}\right) dt + O\left(\frac{1}{\sigma} e^{-\sigma A_j/[(1+\epsilon)m]}\right).$$

Applying Taylor's theorem (to the normal density), we find an approximation correct to $O(\sigma^{-3})$:

$$\frac{\partial P(Y)}{\partial \eta_j} = \frac{(2y_j - 1)}{\sigma}\left\{\phi(-\eta_j/\sigma) \int_{-\infty}^{\infty} h'(t) f_2(t, \sigma^2)dt + \frac{\phi'(-\eta_j/\sigma)}{\sigma} \int_{-\infty}^{\infty} t h'(t) f_2(t, \sigma^2)dt\right\}$$

$$+ O(\sigma^{-3}) + O\left(\frac{1}{\sigma} e^{-\sigma A_j/[(1+\epsilon)m]}\right). \tag{18}$$

We now expand f_2 in terms of $\Delta_{lj} = \eta_l - \eta_j$ to find a computationally simpler expansion. Recall that $I(j) = |\mathcal{S}_1 \cap \mathcal{Z}(j) \setminus \{j\}|$, $J(j) = |\mathcal{S}_0 \cap \mathcal{Z}(j) \setminus \{j\}|$, and note that

$$f_2(t, \sigma^2) = h(t)^{I(j)}\{1 - h(t)\}^{J(j)} + \sum_{l \in \mathcal{S}_1 \cap \mathcal{Z}(j) \setminus \{j\}} \Delta_{lj} h'(t) h(t)^{I(j)-1}\{1 - h(t)\}^{J(j)}$$

$$- \sum_{l \in \mathcal{S}_0 \cap \mathcal{Z}(j) \setminus \{j\}} \Delta_{lj} h'(t) h(t)^{I(j)}\{1 - h(t)\}^{J(j)-1} + \sum_{l \in \mathcal{Z}(j)} O(\Delta_{lj}^2). \tag{19}$$

Substituting (19) into (18) gives the result.

Part (ii). Applying a similar argument to that in the proof of Lemma 1, to the function f_1 in the case $\{\mathcal{S}_1 \cap \mathcal{N}(j)\} \cup \{\mathcal{S}_0 \cap \mathcal{P}(j)\} \ne \emptyset$, shows that

$$\int_{-\infty}^{\infty} h'(t) f_1(t, \sigma^2) f_4(t, \sigma^2)dt = O(e^{-\sigma B_j/(1+\epsilon)}).$$

Applying this to (17) above gives the result. $\square$

Lemma 2. *Let $f_5(\eta, u)$ be a function, measurable as a function of u for fixed η, satisfying $0 \le f_5(\eta, u) \le K$. Then:*

$$\int_{-\infty}^{\infty} f_5(\eta, \sigma u) h(\eta + \sigma u) \phi(u) du = \int_{-\infty}^{\infty} f_5(\eta, \sigma u) S(\eta + \sigma u) \phi(u) du + O(\sigma^{-1}),$$

where $S(t) = \mathbb{I}(t > 0)$.

of Lemma 2. Note that

$$\left| \int_{-\infty}^{\infty} f_5(\eta, \sigma u) \left[h(\eta + \sigma u) - S(\eta + \sigma u) \right] \phi(u) du \right| \le K \frac{1}{\sigma} \int_{-\infty}^{\infty} |h(t) - S(t)| \, \phi(t/\sigma - \eta/\sigma) dt$$

$$\le \frac{K \phi(-\eta/\sigma)}{\sigma} \int_{-\infty}^{\infty} |D(t)| dt + O(\sigma^{-2}),$$

where $D(t) = h(t) - S(t)$, by application of Taylor's theorem. $\qquad \square$

of Theorem 1 (Probabilities). Part (i): Observe

$$P(Y) = \int_{-\infty}^{\infty} \prod_{j \in \mathcal{S}_1} h(\eta_j + \sigma u) \prod_{j \in \mathcal{S}_0} \{1 - h(\eta_j + \sigma u)\} \phi(u) du$$

$$= \int_{-\infty}^{\infty} \prod_{j \in \mathcal{S}_1} \mathbb{I}(\eta_j + \sigma u > 0) \prod_{j \in \mathcal{S}_0} \mathbb{I}(\eta_j + \sigma u < 0) \phi(u) du + O(\sigma^{-1})$$

$$= \int_{-\infty}^{\infty} \mathbb{I}(\max_{j \in \mathcal{S}_0}\{\eta_j/\sigma\} < -u < \min_{j \in \mathcal{S}_1}\{\eta_j/\sigma\}) \phi(u) du + O(\sigma^{-1})$$

$$= \max\{0, \Phi(\min_{j \in \mathcal{S}_1}\{\eta_j/\sigma\}) - \Phi(\max_{j \in \mathcal{S}_0}\{\eta_j/\sigma\})\} + O(\sigma^{-1}),$$

where the second line follows by repeated application of Lemma 2.

Part (ii): By assumption, there exists $j' \in \mathcal{S}$ such that $\{\mathcal{S}_0 \cap \mathcal{P}(j')\} \cup \{\mathcal{S}_1 \cap \mathcal{N}(j')\} = \emptyset$, $|\mathcal{S}_0 \cap \mathcal{Z}(j')| \ge 1$ and $|\mathcal{S}_1 \cap \mathcal{Z}(j')| \ge 1$. Thus, taking $l_1 \in \mathcal{S}_1 \cap \mathcal{Z}(j')$, $l_2 \in \mathcal{S}_0 \cap \mathcal{Z}(j')$,

$$P(Y) = \frac{1}{\sigma} \int_{-\infty}^{\infty} \left[h(\Delta_{l_1 j'} + t)\{1 - h(\Delta_{l_2 j'} + t)\} \right.$$

$$\prod_{l \in \mathcal{S}_0 \cap \mathcal{Z}(j') \setminus \{l_2\}} \{1 - h(\Delta_{l j'} + t)\} \prod_{l \in \mathcal{S}_1 \cap \mathcal{Z}(j') \setminus \{l_1\}} h(\Delta_{l j'} + t)$$

$$\left. f_{3,j'}(t, \sigma^2) \phi(t/\sigma - \eta_{j'}/\sigma) \right] dt.$$

Since $\Delta_{l_1 j'}, \Delta_{l_2, j'} \to 0$, we have that $\Delta_{l_1 j'}, \Delta_{l_2 j'}$ are bounded. Thus, from Lemma 1(ii),

$$P(Y) = \frac{1}{\sigma} \int_{-\infty}^{\infty} \prod_{l \in \mathcal{S}_0 \cap \mathcal{Z}(j')} \{1 - h(\Delta_{l j'} + t)\} \prod_{l \in \mathcal{S}_1 \cap \mathcal{Z}(j')} h(\Delta_{l j'} + t) \phi(t/\sigma - \eta_{j'}/\sigma) dt$$

$$+ O\left(\frac{1}{\sigma} e^{-\sigma A_{j'}/[(1+\epsilon)m]} \right).$$

This can be approximated using a Taylor expansion in $\Delta_{l j'}$ as

$$P(Y) = \frac{1}{\sigma} \int_{-\infty}^{\infty} \prod_{l \in \mathcal{S}_0 \cap \mathcal{Z}(j')} \{1 - h(t)\} \prod_{l \in \mathcal{S}_1 \cap \mathcal{Z}(j')} h(t) \phi(t/\sigma - \eta_{j'}/\sigma) dt$$

$$+ \sum_{l \in \mathcal{Z}(j')} O(\Delta_{l j'}/\sigma) + O\left(\frac{1}{\sigma} e^{-\sigma A_{j'}/[(1+\epsilon)m]} \right).$$

A formal argument using the mean value form of Taylor's theorem can be made to verify that the additional error incurred by the last step is indeed $\sum_{l \in \mathcal{Z}(j')} O(\Delta_{l j'}/\sigma)$. Applying Taylor's theorem to

the normal density function yields

$$P(Y) = \frac{1}{\sigma} \int_{-\infty}^{\infty} \prod_{l \in \mathcal{S}_0 \cap \mathcal{Z}(j')} \{1 - h(t)\} \prod_{l \in \mathcal{S}_1 \cap \mathcal{Z}(j')} h(t)\phi(-\eta_{j'}/\sigma)dt$$

$$+ \frac{1}{\sigma^2} \int_{-\infty}^{\infty} \prod_{l \in \mathcal{S}_0 \cap \mathcal{Z}(j')} \{1 - h(t)\} \prod_{l \in \mathcal{S}_1 \cap \mathcal{Z}(j')} h(t)t\phi'(-\tilde{\eta}_t/\sigma)dt$$

$$+ \sum_{l \in \mathcal{Z}(j')} O(\Delta_{lj'}/\sigma) + O\left(\frac{1}{\sigma}e^{-\sigma A_{j'}/[(1+\epsilon)m]}\right),$$

with $\tilde{\eta}_t$ between $\eta_{j'}$ and $\eta_{j'} - t$. Since $h'(t) = h(t)\{1 - h(t)\}$, the second integral has the form $\int_{-\infty}^{\infty} h'(t)f_4(t, \sigma^2)dt$, with f_4 bounded, and so the overall remainder term is $O(\sigma^{-2})$. $\qquad\square$

of Theorem 3. We show that, for all outcomes,

$$\left|\frac{\partial P(Y)}{\partial \eta_j}\right|/P(Y) \leq 2.$$

Observe that both $h(t), 1 - h(t) \geq (1/2)e^{-|t|}$ and $h'(t) \leq e^{-|t|}$. For $j \in \mathcal{S}_1$,

$$P(Y) = \int_{-\infty}^{\infty} h(\eta_j + \sigma u) \prod_{l \in \mathcal{S}_1 \setminus \{j\}} h(\eta_l + \sigma u) \prod_{l \in \mathcal{S}_0 \setminus \{j\}} \{1 - h(\eta_l + \sigma u)\}\phi(u)du$$

$$\geq (1/2) \int_{-\infty}^{\infty} e^{-|\eta_j + \sigma u|} \prod_{l \in \mathcal{S}_1 \setminus \{j\}} h(\eta_l + \sigma u) \prod_{l \in \mathcal{S}_0 \setminus \{j\}} \{1 - h(\eta_l + \sigma u)\}\phi(u)du,$$

and the same lower bound holds for $j \in \mathcal{S}_0$. Compare with the derivative,

$$\left|\frac{\partial P(Y)}{\partial \eta_j}\right| = \int_{-\infty}^{\infty} h'(\eta_j + \sigma u) \prod_{l \in \mathcal{S}_1 \setminus \{j\}} h(\eta_l + \sigma u) \prod_{l \in \mathcal{S}_0 \setminus \{j\}} \{1 - h(\eta_l + \sigma u)\}\phi(u)du$$

$$\leq \int_{-\infty}^{\infty} e^{-|\eta_j + \sigma u|} \prod_{l \in \mathcal{S}_1 \setminus \{j\}} h(\eta_l + \sigma u) \prod_{l \in \mathcal{S}_0 \setminus \{j\}} \{1 - h(\eta_l + \sigma u)\}\phi(u)du.$$

Thus $\left|\frac{\partial P(Y)}{\partial \eta_j}\right|/P(Y) \leq 2$ and, in conjunction with Theorems 1 and 2, the theorem is proved.

$\qquad\square$

Propositions 1 and 2 below give additional details of the behaviour of the random intercept logistic regression model for large σ^2.

Proposition 1. *As $\sigma^2 \to \infty$ (i) the probability that the outcome in any given block is increasing is $1 + O(\sigma^{-1})$; (ii) the probability that the outcomes in all blocks are increasing is $1 + O(\sigma^{-1})$.*

Proof. Consider a single block. Without loss of generality, we may assume the units in the block are ordered such that $\eta_1 \leq \ldots \leq \eta_m$. We define $\eta_0 = -\infty$, $\eta_{m+1} = \infty$ for convenience. Then, the increasing outcomes are $(00\ldots0)$, $(00\ldots01)$, $(00\ldots11)$, $\ldots$, $(11\ldots1)$. From Theorem 1, with $Y = (y_1, \ldots, y_m)^T \in \{0, 1\}^m$, a within block outcome vector,

$$P\{Y \text{ is increasing and first 1 occurs at } y_j\} = \Phi(-\eta_{j-1}/\sigma) - \Phi(-\eta_j/\sigma) + O(\sigma^{-1}).$$

Overall,

$$P\{Y \text{ is increasing}\} = \sum_{j=1}^{m+1} P\{Y \text{ is increasing and first 1 occurs at } j\text{th position}\}$$

$$= \sum_{j=1}^{m+1} [\Phi(-\eta_{j-1}/\sigma) - \Phi(-\eta_j/\sigma)] + O(\sigma^{-1})$$

$$= \Phi(-\eta_0/\sigma) - \Phi(-\eta_1/\sigma) + \Phi(-\eta_1/\sigma) - \Phi(-\eta_2/\sigma)$$
$$+ \ldots - \Phi(-\eta_m/\sigma) + \Phi(-\eta_m/\sigma) - \Phi(-\eta_{m+1}/\sigma) + O(\sigma^{-1})$$

$$= \Phi(\infty) - \Phi(-\infty) + O(\sigma^{-1})$$

$$= 1 + O(\sigma^{-1}).$$

By independence of blocks, the probability that the outcomes of all blocks are increasing is $(1+O(\sigma^{-1}))^n$. This equals $1 + nO(\sigma^{-1}) + O(\sigma^{-2}) = 1 + O(\sigma^{-1})$, by binomial expansion. $\qquad\square$

Proposition 2. *For any $\sigma > 0$, if all blocks have increasing outcomes, then the parameters of the logistic model with fixed block effects and linear predictor*

$$\eta_{ij} = f^{\mathrm{T}}(x_{ij})\beta + \gamma_i, \quad i = 1, \ldots, n; j = 1, \ldots, m,$$

are not estimable by maximum likelihood.

Proof. The argument is essentially the same as for separation in the standard logistic model case. From the assumptions that the outcomes in each block are increasing, for each i there exists $\tilde{\eta}_i$ such that

$$f^{\mathrm{T}}(x_{ij})\beta > \tilde{\eta}_i \iff y_{ij} = 1$$
$$f^{\mathrm{T}}(x_{ij})\beta < \tilde{\eta}_i \iff y_{ij} = 0$$

For $\lambda > 0$, consider $\theta_\lambda = (\beta_\lambda, \gamma_\lambda) = (\lambda\beta, -\lambda\tilde{\eta})$. Let $\delta_{ij} = f^T(x_{ij})\beta - \tilde{\eta}_i$, and note that $\delta_{ij} > 0$ if $y_{ij} = 1$ and $\delta_{ij} < 0$ if $y_{ij} = 0$. Then

$$\mathrm{Pr}(y|\theta_\lambda) = \prod_{i,j \,:\, y_{ij}=1} h(\lambda\delta_{ij}) \prod_{i,j \,:\, y_{ij}=0} \{1 - h(\lambda\delta_{ij})\}$$

As $\lambda \to \infty$, $\mathrm{Pr}(y|\hat\theta_\lambda) \to 1$. Thus, given any set of finite parameter values (which must have likelihood less than 1), there is a $\hat\theta_\lambda$ that has higher likelihood. Thus there is no set of finite parameter values that maximize the likelihood. $\qquad\square$

Appendix 3: Estimation of parameters for large σ^2

To assess the difficulty of estimating the fixed parameters for varying σ, for parameters $\beta_i \neq 0$ we examined the approximate relative error of estimation,

$$\mathrm{sd}(\hat\beta_i)/|\beta_i| \approx [M_\beta^{-1}(\xi^*; \theta)]_{ii}^{1/2}/(\sqrt{n}|\beta_i|),$$

with the optimal design for each of the parameter combinations in Section 5.2. For $\beta_i = 0$, we compared the standard deviation of $\hat\beta_i$ to the magnitude of the smallest nonzero parameter,

$$\mathrm{sd}(\hat\beta_i)/\min_{\{i:\beta_i\neq 0\}} |\beta_i|.$$

These relative errors are plotted in Figure A1 above, with each colour corresponding to a different parameter scenario. We use relative errors as these are most appropriate when comparing estimation quality for parameter values of potentially quite different sizes.

We see that, for comparable values of the marginal parameters, the relative errors for β_1 and β_2 tend to decrease or remain approximately the same as σ increases. For these parameters, therefore, the same level of estimation precision may be achieved for large σ with no additional experimental units
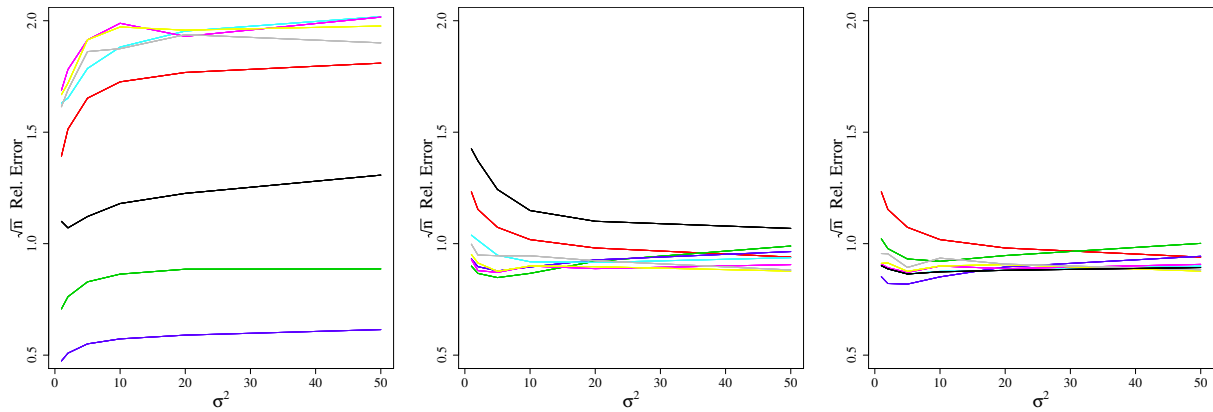
Figure 3: Sample-size normalized approximate relative estimation error (for $\beta_i \neq 0$, $[M_\beta^{-1}(\xi^*; \theta)]_{ii}^{1/2}/\beta_i$), for varying σ^2. Relative errors on the same coloured line correspond to parameter scenarios with the same values of the marginal parameters. The first, second, and third panels correspond to β_0, β_1, and β_2, respectively.

or, in some cases, up to 40% fewer units. The relative error for β_0 increases with σ by 18–30% in our examples. Hence for the largest σ, 39–69% more experimental units are needed to maintain the same level of estimation precision in β_0. However, β_0 is often the parameter of least interest. These sample size considerations make clear that useful experimentation remains possible for large σ though, of course, detailed results for particular applications may vary.

References

Albert, A. & Anderson, J. A. (1984), 'On the existence of maximum likelihood estimates in logistic regression models', *Biometrika* **71**, 1–10.

Breslow, N. E. & Clayton, D. G. (1993), 'Approximate inference in generalized linear mixed models', *J. Am. Statist. Assoc.* **88**(421), 9–25.

Chaganty, N. R. & Joe, H. (2004), 'Efficiency of generalized estimating equations for binary responses', *J. R. Statist. Soc. B* **66**(4), 851–860.

Chaloner, K. & Larntz, K. (1989), 'Optimal Bayesian design applied to logistic regression experiments', *J. Statist. Plan. Infer.* **21**(2), 191–208.

Cheng, C. S. (1995), 'Optimal regression designs under random block-effects models', *Statist. Sinica* **5**, 485–497.

Goldstein, H. & Rasbash, J. (1996), 'Improved approximations for multilevel models with binary responses', *J. R. Statist. Soc. A* **159**(3), 505–513.

Goos, P. & Vandebroek, M. (2001), 'D-optimal response surface designs in the presence of random block effects', *Comp. Statist. & Data Anal.* **37**(4), 433–453.

Joe, H. (1997), *Multivariate Models and Dependence Concepts*, Chapman and Hall/CRC, Boca Raton.

Liang, K. Y. & Zeger, S. L. (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**(1), 13–22.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models*, 2nd edn, Chapman and Hall, London.

Moerbeek, M. & Maas, C. J. M. (2005), 'Optimal experimental designs for multilevel logistic models with two binary predictors', *Commun. Statist. A* **34**(5), 1151–1167.

Müller, P. & Parmigiani, G. (1995), 'Optimal design via curve fitting of Monte Carlo experiments', *J. Am. Statist. Assoc.* **90**(432), 1322–1330.

Niaparast, M. (2009), 'On optimal design for a Poisson regression model with random intercept', *Statist. & Prob. Lett.* **79**(6), 741–747.

Niaparast, M. & Schwabe, R. (2013), 'Optimal design for quasi-likelihood estimation in Poisson regression with random coefficients', *J. Statist. Plan. Infer.* **143**, 296–306.

Nocedal, J. & Wright, S. J. (1999), *Numerical Optimization*, Springer, New York.

Retout, S. & Mentré, F. (2003), 'Further developments of the Fisher information matrix in nonlinear mixed effects models with evaluation in population pharmacokinetics', *J. Biopharma. Statist.* **13**, 209–227.

Russell, K. G., Woods, D. C., Lewis, S. M. & Eccleston, J. A. (2009), 'D-optimal designs for Poisson regression models', *Statist. Sinica* **19**, 721–730.

Santner, T. J., Williams, B. J. & Notz, W. I. (2003), *The Design and Analysis of Computer Experiments*, Springer-Verlag, New York.

Silvey, S. D. (1980), *Optimal Design*, Chapman and Hall, London.

Tekle, F. B., Tan, F. E. S. & Berger, M. P. F. (2008), 'Maximin D-optimal designs for binary longitudinal responses', *Comp. Statist. Data Anal.* **52**(12), 5253–5262.

Waite, T. W. (2013), Integrability and Bayesian D-optimality, Technical report, University of Southampton. http://eprints.soton.ac.uk/id/eprint/355116.

Woods, D. C., Lewis, S. M., Eccleston, J. A. & Russell, K. G. (2006), 'Designs for generalized linear models with several variables and model uncertainty', *Technometrics* **48**(2), 284–292.

Woods, D. C. & Van de Ven, P. (2011), 'Block designs for experiments with correlated non-normal response', *Technometrics* **53**(2), 173–182.

Yang, M., Zhang, B. & Huang, S. (2011), 'Optimal designs for generalized linear models with multiple design variables', *Statist. Sinica* **21**, 1415–1430.

Zeger, S. L., Liang, K. Y. & Albert, P. S. (1988), 'Models for longitudinal data: a generalized estimating equation approach', *Biometrics* **44**(4), 1049–1060.