OPTIMAL DESIGN FOR EXPERIMENTS WITH POSSIBLY INCOMPLETE OBSERVATIONS

Kim May Lee, Stefanie Biedermann and Robin Mitra

University of Southampton, UK

Abstract: Missing responses occur in many industrial or medical experiments, for example in clinical trials where slow acting treatments are assessed. Finding efficient designs for such experiments can be problematic since it is not known at the design stage which observations will be missing. The design literature mainly focuses on assessing robustness of designs for missing data scenarios, rather than finding designs which are optimal in this situation. Imhof, Song and Wong (2002) propose a framework for design search, based on the expected information matrix. We develop a new approach which includes Imhof, Song and Wong (2002)'s method as special case and justifies its use retrospectively. Our method is illustrated through a simulation study based on real data from an Alzheimer's disease trial.

Key words and phrases: Covariance matrix, information matrix, linear regression model, missing observations, optimal design.

1. Introduction

In statistical studies, having missing values in the collected data sets is often unavoidable, in particular when the experimental units are humans and the study is long-term. Consider, for example, a clinical trial where responses are measured several months into the treatment regime for comparison with baseline measurements. In this situation, some patients may be lost to follow-up for various reasons, including side effects of the treatment or death.

Extracting the essential information on treatment characteristics from only partially observed data is a key challenge. Missing values may reduce the power of the study or increase the variability of estimation, due to smaller sample size. Moreover, when not missing completely at random (MCAR), they can cause bias in estimates and thus result in misleading conclusions when not analysed appropriately, see e.g. Little and Rubin (2002), Schafer (1997) or Carpenter, Kenward and White (2007). Several methods have been suggested in the literature to deal with this issue, for example, multiple imputation (Rubin, 1987), maximum likelihood, weighting methods or pattern mixture models. Research in this area has found much attention, see for example Kenward, Molenberghs and Thijs (2003), White, Higgins and Wood (2008) and Spratt, Carpenter, Sterne, Carlin, Heron, Henderson and Tilling (2010).

In this article we assume the missing data problem is handled using a complete case analysis. This approach discards any experimental units containing missing values from the analysis. Usual statistical procedures, such as regression analysis, are then applied to the (reduced) fully observed data set. The approach is appealing because of its simplicity. In addition, inferences of regression coefficients under complete case analysis are unbiased provided the probability responses are missing only depends on the covariates and not on the response itself. The reason for this is because the regression analysis considers the conditional distribution of the responses given the covariates, and so both response and covariates should be present to contribute to the inference. This is a well known result in missing data and has been noted in the literature (Little and Rubin, (2002), Glynn and Laird, (1986))

In the situation of completely observable data, it is well-established that a good design can decrease the necessary sample size, and thus lower the costs of experimentation. However, the design literature so far has only addressed very few special cases involving missing data, which provide only limited guidance to practitioners. Several papers focus on assessing the robustness of standard designs, such as balanced incomplete block designs, D-optimal designs or response surface designs, against missing observations; see, for example, Hedayat and John (1974), Ghosh (1979), Ortega-Azurduy, Tan and Berger (2008) or Ahmad and Gilmour (2010).

Herzberg and Andrews (1976) propose to optimise the expectation of the D- and G-objective functions, respectively, where random missing data indicators are incorporated into the information matrix. Such a modified G-optimal design minimises the expected maximum variance of a predicted response among all designs where these variances exist. Hackl (1995) penalises singular information matrices in a modified version of the D-optimality criterion, and considers only small finite design spaces since the approach would become intractable for continuous intervals or even large discrete sets. Imhof, Song and Wong (2002) develop a framework for finding optimal designs using the expected information matrix, where the expectation is taken with respect to the missing data mechanism. This approach is mathematically equivalent to finding designs for heteroscedastic or weighted regression models. Imhof, Song and Wong (2004) extend this work by exploring different classes of probability functions for missing responses, and study the robustness of their optimal designs against misspecification of the parameters in the probability functions. Baek, Zhu, Wu and Wong (2006) further extend this approach to Bayesian optimality criteria. They study optimal designs for estimating percentiles of a dose-response curve with potentially missing observations.

In the situation where all outcomes will be observed, it is common in the optimal design literature to use the inverse of the information matrix as an approximation to the covariance matrix, $\boldsymbol{var}(\hat{\boldsymbol{\beta}})$, of the parameter estimators of interest, held in the vector $\hat{\boldsymbol{\beta}}$. For linear models, these two matrices are in fact the same. For maximum likelihood estimators in non-linear or generalised linear models, equality holds asymptotically. However, when some of the responses may be missing, $\boldsymbol{var}(\hat{\boldsymbol{\beta}})$ will not exist, and it is not clear if the inverse information matrix will be a good approximation to the observed covariance matrix, i.e. the covariance matrix after the experiment has been carried out. Hence it is not known if a design which is optimal with respect to some function of the expected information matrix will actually make the (observed) covariance matrix (or a function thereof) small. Imhof, Song and Wong (2002) implicitly assumed that this would be the case without providing a justification. Our research is filling this gap. We propose a more sophisticated approximation to the covariance matrix which contains Imhof, Song and Wong (2002)'s method as a special case, and thus justifies their approach retrospectively. The framework proposed in this paper is applicable to finding optimal designs for linear regression models in the presence of missing at random (MAR) mechanisms (or MCAR, which is a special case of MAR).

The structure of the paper is as follows. In Section 2, we provide some background on optimal design for complete data, and describe the optimal design framework for incomplete data proposed by Imhof, Song and Wong (2002). In Section 3, we introduce and justify an optimal design framework for a broad class of MAR missing data mechanisms which includes the method by Imhof, Song and Wong (2002) as a special case. Using a simple linear regression model, the optimal design framework is illustrated for A-, c- and D-optimal designs in Section 4. In Section 5, we apply our framework to redesigning a clinical trial for two Alzheimer's drugs, while providing a discussion of our results in Section 6.

2. Background

We briefly introduce the general linear regression model and some basic theory on optimal design of experiments for the situation where all outcomes are observed. Consider the general linear regression model for $(p+1)$ linearly independent functions $f_0(x), ..., f_p(x)$,

$$Y_i = \beta_0 f_0(x_i) + ... + \beta_p f_p(x_i) + \epsilon_i, \quad x_i \in \mathfrak{X}, \;\; i = 1, \ldots, n, \tag{2.1}$$

where Y_i is the ith value of the response variable, x_i is the value of the explanatory variable (or the vector of explanatory variables) for experimental unit i, $\mathfrak{X}$ is the design region, and $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, $i = 1, \ldots, n$. In matrix form, this can be written as

$$\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$$

where the ith row of $\boldsymbol{X}$ is $\boldsymbol{f}^T(x_i) = (f_0(x_i), \ldots, f_p(x_i))$. A typical example is the polynomial regression model of degree p, i.e.

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_p x_i^p + \epsilon_i. \tag{2.2}$$

Using the method of either least squares or maximum likelihood, the vector of unknown parameters, $\boldsymbol{\beta}$, is estimated by $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{Y}$, with covariance matrix

$$\boldsymbol{var}\ (\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

Let x_i^*, $i = 1, \ldots, m$, $m \leq n$, be the *distinct* values of the explanatory variable in the experimental design, and let n_i, $i = 1, \ldots, m$, be the number of observations taken at x_i where $\sum_{i=1}^m n_i = n$. Then an *exact* design can be written as

$$\xi = \begin{Bmatrix} x_1^* & \cdots & x_m^* \\ w_1 & \cdots & w_m \end{Bmatrix}$$

where $w_i = n_i/n$ gives the proportion of observations to be made in the *support point* x_i^*. This concept can be generalised to *approximate* or *continuous* designs where the restriction that $w_i n$ is a positive integer is relaxed to $w_i > 0$, $i = 1, \ldots, m$, with $\sum_{i=1}^{m} w_i = 1$. The proportion w_i is called the *weight* at the support point x_i^*. The latter approach avoids the problem of discrete optimisation and is widely used in finding optimal designs for experiments. In order to run such a design in practice, a rounding procedure which turns continuous designs into exact designs can be applied; see, for example, Pukelsheim and Rieder (1992). For a continuous design ξ, the Fisher information matrix for model (2.1) is

$$\boldsymbol{M}(\xi) = n \sum_{i=1}^{m} \boldsymbol{f}(x_i^*) \boldsymbol{f}^T(x_i^*) \; w_i$$

and its inverse, $\boldsymbol{M}^{-1}(\xi)$, is proportional to $\boldsymbol{var} \; (\hat{\boldsymbol{\beta}})$.

The design problem is to find the values of x_i^* and w_i that provide maximum information from the experiment. Let Ξ be the class of all possible designs on $\mathfrak{X}$ and $\mathfrak{M}$ be the set of all information matrices with respect to Ξ, i.e. $\mathfrak{M} = \{\boldsymbol{M}(\xi); \xi \in \Xi\}$. An optimality criterion is a statistically meaningful, real-valued function $\psi(\boldsymbol{M}(\xi))$, which is selected to reflect the objective of the experiment. It is typically an increasing and convex function over $\mathfrak{M}$, such that there is a critical point in the region. The technical explanation of these properties can be found in experimental design books such as Silvey (1980) and Pukelsheim (2006). We seek a design ξ^* such that $\psi(\boldsymbol{M}(\xi^*)) = \min\limits_{\xi \in \Xi} \psi(\boldsymbol{M}(\xi))$. Such a design is called a ψ-optimal design.

The following optimality criteria are some examples commonly used in finding the optimal setting for an experiment with the corresponding objective.

- D-optimality: $\psi(\boldsymbol{M}(\xi)) = |\boldsymbol{M}^{-1}(\xi)|$. A D-optimal design minimises the volume of a confidence ellipsoid for $\boldsymbol{\beta}$.

- A-optimality: $\psi(\boldsymbol{M}(\xi)) = \text{trace}(\boldsymbol{M}^{-1}(\xi))$. An A-optimal design minimises the sum of the variances of the individual elements of $\hat{\boldsymbol{\beta}}$.

- c-optimality: $\psi(\boldsymbol{M}(\xi)) = \boldsymbol{c}^T \boldsymbol{M}^{-1}(\xi) \boldsymbol{c}$ where $\boldsymbol{c}$ is a $(p+1) \times 1$ vector. A c-optimal design minimises the variance of $\boldsymbol{c}^T \hat{\boldsymbol{\beta}}$, a given linear combination of $\hat{\boldsymbol{\beta}}$.

2.1 Optimal design for missing values

To construct optimal designs that account for missing observations, we define independent random missing data indicators $R_i = 1$, if the observation at x_i is missing; $R_i = 0$ otherwise, $i = 1, \ldots, n$. Following Rubin (1976), if responses are missing completely at random (MCAR) then

$$Pr(R_i = 1 | x_i, y_i, i = 1, ..., n) = P(R_i) \qquad \forall i = 1, \ldots, n.$$

If we have a missing at random (MAR) mechanism the probability of missingness may depend on the observed values of x_i and y_i, i.e. for $i = 1, \ldots, n$,

$$Pr(R_i = 1 \mid x_i, y_i, i = 1, ..., n) = E\{R_i \mid \text{observed } x_i, y_i, i = 1, ..., n\}.$$

In what follows, since only the design values of x_i play a role in the optimal design framework, we assume a special case of MAR mechanism where

$$E\{R_i \mid \text{observed } x_i, y_i, i = 1, ..., n\} = P(R_i = 1 \mid \text{observed } x_i) = P(x_i).$$

This is necessary as we do not know which responses will be observed at the time of designing the experiment. In the remaining part of this paper, the conditioning on x_i will be omitted to simplify the notation of a MAR mechanism.

The Fisher information matrix containing the missing data indicators $\boldsymbol{R} = \{R_1, R_2, \ldots, R_n\}$ is given by

$$\begin{aligned} \boldsymbol{E}\{\boldsymbol{M}(\xi, \boldsymbol{R})\} &= \boldsymbol{E}\{\sum_{i=1}^{n} \boldsymbol{f}(x_i)\boldsymbol{f}^T(x_i)\ (1 - R_i)\} \\ &= \sum_{i=1}^{n} \boldsymbol{f}(x_i)\boldsymbol{f}^T(x_i)\ (1 - P(x_i)) \\ &= n\sum_{i=1}^{m} \boldsymbol{f}(x_i^*)\boldsymbol{f}^T(x_i^*)\ w_i\ (1 - P(x_i^*)) \qquad (2.3) \end{aligned}$$

which is equivalent to $\boldsymbol{M}(\xi)$ if the responses are fully observed.

Imhof, Song and Wong (2002) proposed a general framework where a function of (2.3) is used in constructing optimal designs. For example, a D-optimal design maximises $|\boldsymbol{E}\{\boldsymbol{M}(\xi, \boldsymbol{R})\}|$ as $\boldsymbol{var(\hat{\beta})}$ was implicitly assumed to be proportional to $[\boldsymbol{E}\{\boldsymbol{M}(\xi, \boldsymbol{R})\}]^{-1}$.

The use of $\boldsymbol{E}\{\boldsymbol{M}(\xi, \boldsymbol{R})\}$ is appealing since $\boldsymbol{M}(\xi, \boldsymbol{R})$ is linear in the missing data indicators, and therefore taking the expectation is straightforward. Moreover, from (2.3), we can see that this framework is analogous to the optimal design framework for weighted regression models, with weight function $\lambda(x) = 1 - P(x)$.

However, if responses may be missing, $\boldsymbol{var(\hat{\beta})}$ does not exist. Hence it is not clear if the inverse of $\boldsymbol{E}\{\boldsymbol{M}(\xi, \boldsymbol{R})\}$ will be a good approximation to the observed covariance matrix of an experiment. In the next section, we will investigate this approximation further.

3. Optimal design for MAR mechanisms with complete case analysis

We assume that the missing data mechanism is MAR and consider the law of total variance,

$$\boldsymbol{var}(\hat{\boldsymbol{\beta}}) = \boldsymbol{E}(\boldsymbol{var}(\hat{\boldsymbol{\beta}}|\boldsymbol{R})) + \boldsymbol{var}(\boldsymbol{E}(\hat{\boldsymbol{\beta}}|\boldsymbol{R})). \tag{3.1}$$

This law only holds if all terms exist, but existence is not guaranteed due to the presence of missing data. It is possible that the values of $\boldsymbol{R}$ result in $\boldsymbol{X}^T\boldsymbol{X}$ being singular (for example if all values $R_i = 1$) but in the situations we consider such occurrences would be extremely rare. Thus, if v is the probability that $\boldsymbol{X}^T\boldsymbol{X}$ is singular due to $\boldsymbol{R}$, we assume that v is negligibly small, and to allow us to apply (3.1) above we assume that in such situations we assign some "dummy value" to $\hat{\boldsymbol{\beta}}$, for example an estimate based on some prior distribution from a Bayesian approach, although in principle any value would do. Whatever value is chosen will only have a negligibly small influence on the final result when applying (3.1) due to the very small value of v.

Let $\mathcal{C}$ be the set of values of $\boldsymbol{R}$ such that $\boldsymbol{X}^T\boldsymbol{X}$ is non-singular. Then $\boldsymbol{E}(\hat{\boldsymbol{\beta}}|\boldsymbol{R} \in \mathcal{C}) = \boldsymbol{\beta}$ and $\boldsymbol{var}(\boldsymbol{E}(\hat{\boldsymbol{\beta}}|\boldsymbol{R} \in \mathcal{C})) = \boldsymbol{var}(\boldsymbol{\beta}) = \boldsymbol{0}$, the $(p+1) \times (p+1)$ matrix with all elements equal to zero. Since v is assumed to be negligibly small we approximate the covariance matrix by

$$\boldsymbol{var}(\hat{\boldsymbol{\beta}}) \approx \boldsymbol{E}(\boldsymbol{var}(\hat{\boldsymbol{\beta}}|\boldsymbol{R})).$$

Again, using the fact that v is close to zero, the right hand side should be close to $\boldsymbol{E}\{[\boldsymbol{M}(\xi, \boldsymbol{R})^{-1}]\}$. We apply a multivariate second-order Taylor series expansion to approximate the elements of the inverse matrix $\boldsymbol{M}(\xi, \boldsymbol{R})^{-1}$, in the hope that the approximated value of $\boldsymbol{E}\{[\boldsymbol{M}(\xi, \boldsymbol{R})^{-1}]\}$ is close to the observed value of $\boldsymbol{var(\hat{\beta})}$. The approach by Imhof, Song and Wong (2002) can be viewed as a Taylor expansion of order one, where they implicitly approximate $\boldsymbol{E}\{[\boldsymbol{M}(\xi,\boldsymbol{R})^{-1}]\}$ by $[\boldsymbol{E}\{\boldsymbol{M}(\xi,\boldsymbol{R})\}]^{-1}$. Technically the order of the approximation could be viewed as either the 0th or 1st order. While no Taylor expansion has actually been applied here, it could be viewed as the 0th order expansion, but as we are expanding the expression about the mean of the random variables, the first order expansion simplifies to the 0th order result. As our approach is obtained using a second Taylor expansion about the mean, we refer to the Imhof et. al approach as the 1st order approach for consistency.

While the first order expansion will usually provide a cruder approximation to the 'true' objective function, and thus somewhat less efficient designs, this approach has the advantage that established theory on optimal design, such as the use of equivalence theorems, is applicable. Hence we can often simplify design search considerably through analytical results. For second order approximations, convexity of the objective function is no longer guaranteed, which prohibits the use of equivalence theorems. Hence, while optimal designs will be more efficient, analytical results can only be established on a case by case basis, and design search will be more challenging.

Theorem 1 shows that for a large class of MAR mechanisms and polynomial models, the D-optimal design found using a first order approximation has the same number of support points as it has parameters. This result corresponds to the contribution of De la Garza (1954) and Silvey (1980) in the conventional optimal design framework for finding the number and weight of support points of a D-optimal design. The proof of Theorem 1 can be found in Appendix A.1.

Theorem 1. *Let $h(x) = \frac{1}{1-P(x)}$ and assume that for the MAR mechanism $P(x)$ the equation $h^{(2p)}(x) = c$ has at most one solution for every constant $c \in \Re$. Then a D-optimal design for the polynomial model (2.2) of degree p has exactly $p+1$ support points, with equal weights.*

Hence design search can be restricted to $(p+1)$-point designs, with known weights $w_i = 1/(p+1)$, $i = 1, \ldots, p+1$. A further simplification is given in Lemma (2), which shows that under the assumptions of Theorem 1, if the MAR mechanism is monotone, one of the bounds of the design region is a support point of the D-optimal design.

Lemma 2. *Let $P(x)$ be a MAR mechanism that satisfies the conditions in Theorem 1 and is monotone, and let the design interval $\mathfrak{X} = [l, u]$, where $l < u$. If $P(x)$ is strictly increasing, then the lower bound, l, is a support point of the D-optimal design. If $P(x)$ is strictly decreasing, then the upper bound, u, is a support point of the D-optimal design.*

Proof. For a continuous design ξ with $p+1$ support points, we have

$$|\boldsymbol{E}\{\boldsymbol{M}(\xi, \boldsymbol{R})\}| \;=\; \prod_{i=1}^{p+1} w_i(1 - P(x_i^*)) \prod_{1 \le i < j \le p+1} (x_i^* - x_j^*)^2 \tag{3.2}$$

where we order the support points by size:

$$l \le x_1^* < x_2^* < ... < x_{p+1}^* \le u.$$

If $P(x)$ is monotonic increasing in x, $(1 - P(x))$ will be largest at $x_1^* = l$ and $(x_1^* - x_j^*)^2$ will also be largest for $x_1^* = l$, for all values of x_j^* where $j = 2, \ldots, p+1$. Hence l must be a support point. Analogously, if $P(x)$ is monotonic decreasing, $(1 - P(x))$ and $(x_i^* - x_{p+1}^*)^2$, $i = 1, \ldots, p$ will be maximised at $x_{p+1}^* = u$. □

For optimal designs based on a second order approximation to $\boldsymbol{E}\{[\boldsymbol{M}(\xi, \boldsymbol{R})^{-1}]\}$, there is no corresponding result in general. However, in the following section, we provide a similar result for a special case.

3.1. Illustration

To fix ideas, we consider the simple linear regression model, i.e. model (2.2) where $p = 1$, for D-, c- and A-optimality. For a design region $\mathfrak{X} = [l, u]$ where $l < u$, consider total sample size n and two support points x_1^* and x_2^*. Two support points are sufficient for estimation in the simple linear regression model with two unknown parameters and, from Theorem 1, the D-optimal designs based on the first order approximation are two-point designs for a large variety of MAR mechanisms $P(x)$. Hence finding the best two-point design for the second order approximation facilitates comparing the two approaches. Let $n_1 = nw_1$ responses $\{y_1,\ ...,\ y_{n_1}\}$ be taken at experimental condition x_1^*, and $n_2 = n - n_1 = nw_2$ responses $\{y_{n_1+1},\ ...,\ y_n\}$ at x_2^*. We seek an optimal design

$$\xi^* = \begin{Bmatrix} x_1^* & x_2^* \\ w_1 & w_2 \end{Bmatrix}$$

based on a function of the approximated expression for $\boldsymbol{E}\{[\boldsymbol{M}(\xi, \boldsymbol{R})^{-1}]\}$. For the simple linear regression model,

$$\boldsymbol{M}(\xi, \boldsymbol{R})^{-1} = \frac{1}{(x_1^* - x_2^*)^2 \, Z_1 Z_2} \begin{pmatrix} x_1^{*2} Z_1 + x_2^{*2} Z_2 & -x_1^* Z_1 - x_2^* Z_2 \\ -x_1^* Z_1 - x_2^* Z_2 & Z_1 + Z_2 \end{pmatrix}, \tag{3.3}$$

where $Z_1 = \sum_{i=1}^{n_1}(1-R_i)$ and $Z_2 = \sum_{i=n_1+1}^{n}(1-R_i)$ follow binomial distributions with parameters $(nw_1,\ 1 - P(x_1^*))$ and $(nw_2,\ 1 - P(x_2^*))$ respectively. If all observations at a support point are missing, i.e. $Z_1 = 0$ or $Z_2 = 0$, $\boldsymbol{M}(\xi, \boldsymbol{R})$ becomes singular and we cannot estimate the model parameters. In the scenarios we investigate, the chance of this occurring is assumed to be quite small. If this possibility is not small, then some changes to the experiment would need to be proposed, e.g. increasing the sample size.

In what follows, we assume that this probability is close to zero. We aim to approximate

$$\boldsymbol{E}\{[\boldsymbol{M}(\xi, \boldsymbol{R})^{-1}]\} = \frac{1}{(x_1^* - x_2^*)^2} \begin{pmatrix} x_1^{*2} E\left(\frac{Z_1}{Z_1 Z_2}\right) + x_2^{*2} E\left(\frac{Z_2}{Z_1 Z_2}\right) & -x_1^* E\left(\frac{Z_1}{Z_1 Z_2}\right) - x_2^* E\left(\frac{Z_2}{Z_1 Z_2}\right) \\ -x_1^* E\left(\frac{Z_1}{Z_1 Z_2}\right) - x_2^* E\left(\frac{Z_2}{Z_1 Z_2}\right) & E\left(\frac{Z_1}{Z_1 Z_2}\right) + E\left(\frac{Z_2}{Z_1 Z_2}\right) \end{pmatrix}. \tag{3.4}$$

due to the fact that the distribution of $\frac{Z_i}{Z_i Z_j}$ is intractable. Using a multivariate second order Taylor series approximation expanded about $E\{Z_i\}$ and $E\{Z_i Z_j\} = E\{Z_i\}E\{Z_j\}$, and taking expectation with respect to the binomial random variables, we obtain

$$E\left(\frac{Z_i}{Z_i Z_j}\right) \approx \frac{1}{nw_j(1 - P(x_j^*))} + \frac{P(x_i^*)P(x_j^*)}{nw_i(1 - P(x_i^*))(nw_j(1 - P(x_j^*)))^2} + \frac{P(x_j^*)}{(nw_j(1 - P(x_j^*)))^2} \tag{3.5}$$

for $i, j = 1, 2,\ i \neq j$. A full derivation of this result is given in Appendix A.2. If the missing data mechanism is MCAR, this expression simplifies to

$$\begin{aligned} E\left(\frac{Z_i}{Z_i Z_j}\right) &\approx \frac{1}{nw_j(1-P)} + \frac{P^2}{nw_i(1-P)(nw_j(1-P))^2} + \frac{P}{(nw_j(1-P))^2} \\ &= \frac{1}{n'w_j} + \frac{P^2}{n'w_i(n'w_j)^2} + \frac{P}{(n'w_j)^2} \end{aligned} \tag{3.6}$$

independent of the values of the support points, where $P = P(R_i)$ is the probability that a response is missing completely at random and $n' = n(1 - P)$ corresponds to a reduced total sample size of the experiment.

We further note that $E\left(\frac{Z_i}{Z_iZ_j}\right) \neq E\left(\frac{1}{Z_j}\right)$ as the probability that $Z_i = 0$, while typically small, is strictly greater than 0. Hence cancelling Z_i in the above expression (which would lead to a slightly simpler expression for the Taylor expansion) would add another level of approximation to the objective function.

After selecting a specific missing data mechanism $P(x)$, the optimal design ξ^* can be found by taking derivatives of the criterion with respect to the support points and weights respectively, with constraints $w_1 + w_2 = 1$ and $x_2^* > x_1^* \in \mathfrak{X}$. Note that in order to define the quantities in (3.3) and below, we need to work in terms of exact designs, i.e. $n_1 = nw_1$ and $n_2 = nw_2$ are integers. To facilitate the numerical computation of the optimal designs, we only use the constraint $w_1 + w_2 = 1$ and then round nw_1^* and nw_2^* to the nearest integers, where w_1^* and w_2^* are the resulting optimal weights. For example, a D-optimal design minimises the determinant of (3.4), i.e

$$\frac{1}{(x_1^* - x_2^*)^2} E\left(\frac{Z_1}{Z_1Z_2}\right) E\left(\frac{Z_2}{Z_1Z_2}\right) \tag{3.7}$$

over $\mathfrak{X}$; a c-optimal design for minimising the variance of $\hat{\beta}_1$, i.e. where $\boldsymbol{c} = (0 \;\; 1)^T$, minimises

$$\frac{1}{(x_1^* - x_2^*)^2} \left(E\left(\frac{Z_1}{Z_1Z_2}\right) + E\left(\frac{Z_2}{Z_1Z_2}\right)\right) \tag{3.8}$$

over $\mathfrak{X}$; an A-optimal design minimises

$$\frac{1}{(x_1^* - x_2^*)^2} \left((x_1^{*2} + 1) E\left(\frac{Z_1}{Z_1Z_2}\right) + (x_2^{*2} + 1) E\left(\frac{Z_2}{Z_1Z_2}\right)\right) \tag{3.9}$$

over $\mathfrak{X}$, where the expectations are approximated by (3.5) or (3.6), depending on the form of the missing data mechanism.

Theorem 3 shows that the D, c- and A-optimal two-point designs based on the second order expansion have a similar structure to the corresponding first order designs. Here the c-optimal design minimises the variance of the estimated slope parameter of the simple linear model.

Theorem 3. *For the simple linear regression model (2.2) with $p = 1$, assume we approximate $\boldsymbol{E}\{[\boldsymbol{M}(\xi, \boldsymbol{R})^{-1}]\}$ by a second order Taylor expansion, and let the design interval $\mathfrak{X} = [l, u]$.*

(a) If the missing data mechanism is MCAR, then the D- and the c-optimal design among the two-point designs are equally weighted on l and u. If, in addition, $l \geq 0$ or $u \leq 0$, the two-point A-optimal design will also have support points l and u.

(b) If the missing data mechanism is MAR and monotone increasing (decreasing), then l (u) is a support point of the D- and the c-optimal design among the two-point designs. If, in addition, $l \geq 0$ ($u \leq 0$), this result also holds for A-optimality among the two-point designs.

The proof of part (a) of Theorem 3 is given in Appendix A.3.

Proof of Theorem 3 (b). Let without loss of generality $x_1^* < x_2^*$, and assume $P(x)$ is monotone increasing. From (3.5), it can be seen that the second order approximations for $E[Z_1/(Z_1Z_2)]$ and for $E[Z_2/(Z_1Z_2)]$ are both increasing in x_1^* and are hence minimised when $x_1^* = l$. Since $x_1^* = l$ also minimises $1/(x_1^* - x_2^*)^2$, and all expressions are non-negative, the objective functions in (3.7) and (3.8) are both minimised when $x_1^* = l$. If $l \geq 0$, $(x_1^{*2} + 1)$ is also increasing in x_1^*, and the result for A-optimality follows.

An analogous argument shows that $x_2^* = u$ minimises (3.7), (3.8) and, for $u \leq 0$, also (3.9) if $P(x)$ is monotone decreasing. □

From part (a), we see that the optimal designs are the same as for the simple linear regression model without missing data. In part (b), we find that the lower/upper limit of the design interval is a support point, and thus has the same support structure as the first order design from Lemma 2. However, the weights and the other support point may have different values. In particular, second order D-optimal designs are not necessarily equally weighted.

In the next section, we find some optimal designs for the two respective approximation strategies and illustrate their performance through simulations.

4. Simulation study

We set the design region $\mathfrak{X} = [0, 2]$ and sample size $n = 30$. For a given design we simulate a response variable by

$$Y_i = 1 + x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

where we treat σ^2 as known. We then introduce missing values into the observed y_i, $i = 1, \ldots, n$, by specifying a MAR mechanism through the following logistic model,

$$P(x_i) = \frac{exp(\gamma_0 + \gamma_1 x_i)}{1 + exp(\gamma_0 + \gamma_1 x_i)}$$

with $\gamma_0 = -4.572$ and $\gamma_1 = 3.191$. The positive value of γ_1 indicates the mechanism is monotone increasing with x_i. The logistic model is a commonly used choice for modelling the missing data mechanism (Ibrahim and Lipsitz (1999), Bang and Robins (2005), Mitra and Reiter (2011, 2016)) as in practical situations, it allows the estimation of parameters in the missing data model using a logistic regression. However, we note there are many other choices for modelling the missing data mechanism (Little (1995)) and our approach would be compatible with any choice of missing data model. We assume the analyst will fit a simple linear regression model to the complete case data, obtaining estimates of the coefficients, $(\hat{\beta}_0, \hat{\beta}_1)$, and their variances, from the available cases, i.e. using only those units for which y_i is observed.

Using Theorem 1 and Lemma 2, the lower bound of $\mathfrak{X}$, 0, is chosen as one of the support points of the two-point optimal design, denoted by x_1^* here. Substituting the MAR mechanism, the value of x_1^* and $w_1 = 1 - w_2$ into the corresponding elements of (3.4), an optimal design is found by minimising a function of this matrix with respect to x_2^* and w_2 in *Mathematica* with the *Minimize* function.

We first consider several designs $\xi = \{0, x_2^*\}$ and, under each design, compare the two proposed approaches for approximating elements of the matrix specified in (3.4), as well as various relevant functions of this matrix. The two proposed approaches approximate $E\left(\frac{Z_i}{Z_i Z_j}\right)$ in the elements of (3.4) with first order and second order Taylor expansions respectively, where the first order expansion corresponds to the approach proposed by Imhof, Song and Wong (2002). Specifically for each design, we repeatedly simulate incomplete data using the models described above and empirically obtain the estimates for (3.4) by averaging the elements in $\boldsymbol{M}(\xi, \boldsymbol{R})^{-1}$, given in (3.3), across the replications. Treating these as the true values, we can then compare the two approximations. Table 4.1 presents the simulation results over 200000 replications from two different designs where $x_2^* = 1$ and $x_2^* = 1.5$ respectively. Consider the design where $x_2^* = 1.5$, we can see that for the $[2, 2]$ element in (3.4) which is the criterion used in c-optimality to minimise the variance of $\hat{\beta}_1$ the first order approximation has a bias of 7.2%, while for the second order approximation this bias has reduced to 1.9%. For this same design, the trace of matrix (3.4) (which corresponds to the criterion used in A-optimality) has a bias of 4.4% and the determinant of the matrix (which corresponds to the criterion used in D-optimality) has a bias of 10.1% when using the first order approximation, while the biases are reduced to 1.1% and 2.6% respectively when using the second order approximation. In general, we can see that using the second approximation yields better approximations of the elements of (3.4) and relevant functions of the matrix.

Table 4.1: Simulation output of 200000 replications for two different designs with weight $w_1 = 0.5 = w_2$, $P(x_1^*) = P(0) = 0.01$, and $n = 30$. The numbers in the penultimate row indicate the frequency of the cases where $\boldsymbol{M}(\xi, \boldsymbol{R})$ becomes singular.

ξ	$\{0, 1\}$	$\{0, 1.5\}$
$[1, 1]$ element of (3.4)	0.06740	0.06740
First order Taylor series approximation	0.06736	0.06736
Second order Taylor series approximation	0.06740	0.06741
$[2, 2]$ element of (3.4)	0.15242	0.10375
First order Taylor series approximation	0.15078	0.09628
Second order Taylor series approximation	0.15222	0.10179
$[1, 2]$ element of (3.4)	-0.06740	-0.04494
First order Taylor series approximation	-0.06736	-0.04490
Second order Taylor series approximation	-0.06740	-0.04494
Determinant of (3.4)	0.00573	0.00497
First order Taylor series approximation	0.00562	0.00447
Second order Taylor series approximation	0.00572	0.00484
No. of cases failed	0	23
$P(x_2^*)$	0.20085	0.55342

We now find optimal values for x_2^* and w_2, over the design region $\mathfrak{X} = [0, u]$ with $w_1 = 1 - w_2$, and where the missing mechanism is defined as above. Table 4.2 present the optimal values when constructing A-optimal, c-optimal and D-optimal designs respectively. We can see that using the 2nd order approximations results in an upper design point that is smaller than the upper design point when using the first order approximation. The final row in the table considers the probability, $P(singular)$, that the regression coefficients cannot be estimated, i.e. the covariance matrix becomes singular. This is equivalent to finding the probability that all outcomes at either one (or both) of the design points are completely missing. For more complicated scenarios, this probability can be calculated as follows (see Imhof et al., 2002):

$$\begin{aligned} P(singular) &= \sum_{j=0}^{m-1} \sum_{\substack{S \subset \{1,\ldots,k\} \\ |S|=j}} P(n_i > 0 \text{ if } i \in S; n_i = 0 \text{ if } i \notin S) \\ &= \sum_{j=0}^{m-1} \sum_{\substack{S \subset \{1,\ldots,k\} \\ |S|=j}} \prod_{i \in S} \left[1 - p(x_i)^{Nw_i}\right] \prod_{i \notin S} p(x_i)^{Nw_i}. \end{aligned}$$

We see that this probability is consistently smaller when adopting the second order approximation over the first order. We also additionally consider a design that assumes the data will be fully observed and places half the observations at both end of the design space, here assumed to be $[0, 2]$. We see that $P(singular)$ is significantly higher here than for other designs, and is motivation for considering the potential for missing data at the design stage of an experiment.

Table 4.2: Optimal designs found by using a first order and a second order Taylor series approximation to (3.4) respectively, for the optimality criterion denoted by the subscript, for $n = 30$. Missing values are introduced through the logistic model with $\gamma_0 = -4.572$ and $\gamma_1 = 3.191$. The other support point is $x_1^* = 0$ with $w_1 = 1 - w_2$ and $P(x_1^*) = 0.01$. ξ is the A, c, and D-optimal design that assumes fully observed responses.

	$\xi^*_{A\ 2nd}$	$\xi^*_{A\ 1st}$	$\xi^*_{c\ 2nd}$	$\xi^*_{c\ 1st}$	$\xi^*_{d\ 2nd}$	$\xi^*_{d\ 1st}$	ξ
x_2^*	1.46206	1.51466	1.54924	1.60059	1.33597	1.37660	2
w_2	0.4665	0.4539	0.6257	0.6208	0.5110	0.5	0.5
$P(x_2^*)$	0.5233	0.5650	0.5919	0.6308	0.4234	0.4553	0.8594
$P(singular)$	1.15 e-04	3.378 e-04	4.7067 e-05	0.0001577	2.5192 e-06	7.4897 e-06	0.10302

To investigate the issue of possible singularity of the covariance matrix further, we consider the effect of varying the parameter values for the missing data mechanism, resulting in different probabilities of missingness at the design points. We focus on D-optimality, and on comparing the designs found using a first order and a second order Taylor series approximation to (3.4), respectively.

Table 4.3 shows some examples of $P(singular)$ computed using the D-optimal designs for the simple linear model found for the different approximation methods with logistic MAR mechanisms. We find that as the probability that a response being missing increases (i.e. γ_0 becomes larger), the optimal designs

Table 4.3: Probability of obtaining a singular covariance matrix using D-optimal designs found using different approximations. The MAR mechanism follows the logistic model with $\gamma_1 = 3.191$; $N = 30$; $x_1 = 0$ and $w_1 = 1 - w_2$.

	$2nd$ order D-optimal design			$1st$ order D-optimal design		
γ_0	x_2^*	w_2	$P(singular)$	x_2^*	w_2	$P(singular)$
-4.572	1.3360	0.5110	2.519 e-06	1.3766	0.5	7.490 e-06
-1.572	0.7554	0.5263	0.003185	0.8362	0.5	0.01325
-0.572	0.6159	0.5353	0.02879	0.7336	0.5	0.09426

found by the first order approach have a consistently higher failure rate in estimating the model parameters.

To further illustrate performance, for each design given in Table 4.2 we repeatedly simulate the incomplete data 200000 times as described above, setting $\sigma^2 = 1$. In each incomplete data set, we compute the sample estimate $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ from the complete cases, i.e. where the design matrix $\boldsymbol{X}$ and response vector $\boldsymbol{y}$ comprises only units with observed response values. We can then empirically obtain the covariance matrix for $\hat{\boldsymbol{\beta}}$ across the replications. Table 4.4 summarises the performance of the designs derived under the different optimality criteria and approximations. We see that the designs obtained under A-optimality have the smallest trace of the covariance matrix for $\hat{\boldsymbol{\beta}}$, as expected. Further, this trace is smaller when using the design obtained from the second order approximation rather than the first order approximation. This pattern is repeated for the other optimality criteria. The design obtained under c-optimality from the 2nd order approximation results in the smallest variance for $\hat{\beta}_1$, and the design obtained under D-optimality from the 2nd order approximation results in the smallest determinant of the covariance matrix for $\hat{\boldsymbol{\beta}}$. The design that assumes fully observed outcomes performs the worst across all optimality criteria, it also has the greatest proportion of cases where it was not possibly to estimate the regression coefficients, as expected. This is further motivation for considering the potential for missing data at the design stage, to extract the most information out of an experiment. In addition, we also note that the second order approximation consistently resulted in fewer cases where it was not possible to estimate the parameters due to the missing data, and reflects what is seen in Table 4.2. This is further motivation for adopting the 2nd order approximation over the 1st order here.

Table 4.4: Simulation outputs of 200000 replications for different designs. The numbers in the last row indicate the frequency of the cases where $\boldsymbol{M}(\xi, \boldsymbol{R})$ becomes singular.

	sample $var(\hat{\beta}_1)$	tr(sample $\boldsymbol{var}(\hat{\boldsymbol{\beta}})$)	|sample $\boldsymbol{var}(\hat{\boldsymbol{\beta}})$|	No. of cases failed
$\xi^*_{A\ 2nd}$	1.0687e-01	**1.6988e-01**	4.8764e-03	19
$\xi^*_{A\ 1st}$	1.0823e-01	1.7123e-01	5.0880e-03	67
$\xi^*_{c\ 2nd}$	**9.7349e-02**	1.8893e-01	5.4165e-03	16
$\xi^*_{c\ 1st}$	9.8102e-02	1.8968e-01	5.7121e-03	35
$\xi^*_{d\ 2nd}$	1.0401e-01	1.7590e-01	**4.5809e-03**	0
$\xi^*_{d\ 1st}$	1.0486e-01	1.7197e-01	4.6526e-03	2
ξ	1.4029e-01	2.0063e-01	7.5657 e-03	20588

We have empirically evaluated the framework proposed to construct optimal designs in the presence of missing values under the assumption of a complete case analysis. We have seen that this framework worked well in the simulations, with evidence suggesting that the second order approximation, at least in these simulations, had the potential to provide better approximations and hence result in better designs. Moreover, in all scenarios we investigated, the probability of a singular covariance matrix was lowest for the optimal design using the second order approximation. In the next section we consider a scenario motivated from an application concerned with designing a clinical trial to treat Alzheimer's disease.

5. Application: Redesigning a study on Alzheimer's disease

To illustrate an application of our approach, we use data from an Alzheimer's disease study which investigated the benefits of administering the treatments donepezil, memantine, and the combination of the two, to patients over a period of 52 weeks, on various quality of life measures. See Howard et al. (2012) for full details of the study. The total number of patients included in the primary intention-to-treat sample was 291, with 72 in the placebo group (Group 1), 74 in the memantine treatment group (Group 2), 73 in the donepezil treatment group (Group 3), and 72 in the donepezil-memantine group (Group 4).

In the per-protocol analysis, 43 patients were excluded in Group 1, 32 in Group 2, 23 in Group 3 and 21 in Group 4. Considering these patients as data missing at random, a logistic regression model is fitted to the data, specifically

$$P(R_i = 1|x_i, v_i) = \frac{exp(\gamma_0 + \gamma_1 x_i + \gamma_2 v_i)}{1 + exp(\gamma_0 + \gamma_1 x_i + \gamma_2 v_i)}$$

where $x_i, v_i \in \{0, 1\}$ represent the level of donepezil and memantine respectively (with 1 indicating the treatment is applied) for patient i. From the data the regression coefficients were estimated to be $\hat{\gamma}_0 = 0.2636472$, $\hat{\gamma}_1 = -0.8988845$ and $\hat{\gamma}_2 = -0.4108504$.

We assume a linear regression model will be fit to the data, i.e.

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 v_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n, \tag{5.1}$$

where Y_i corresponds to the outcome value for patient i. We assume σ^2 is known and fixed to 1 without loss of generality. The specific values of $\beta_0, \beta_1, \beta_2$ will not affect the performance of the different designs. We can define the four groups (G_1 - G_4) the units are allocated to in terms of the design variables x and v:

- G_1: $x_i^* = 0, v_i^* = 0$ with n_1 experimental units;
- G_2: $x_i^* = 0, v_i^* = 1$ with n_2 experimental units;
- G_3: $x_i^* = 1, v_i^* = 0$ with n_3 experimental units;
- G_4: $x_i^* = 1, v_i^* = 1$ with n_4 experimental units.

In this situation we have thus fixed the design points, defined by the values of (x, v) and equal to $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$. The design problem is then to find the optimal number of patients to allocate to Groups G_1 - G_4, denoted by n_1, n_2, n_3, and n_4 respectively, under the assumption the analyst fits a linear regression model of the form described in (5.1) using the complete cases. The diagonal elements of the (conditional) covariance matrix of the least squares estimators for this model are

$$var(\hat{\beta}_0|\boldsymbol{R}) = \frac{Z_2Z_3 + Z_2Z_4 + Z_4Z_3}{Z_1Z_2Z_3 + Z_1Z_2Z_4 + Z_1Z_3Z_4 + Z_2Z_3Z_4},$$

$$var(\hat{\beta}_1|\boldsymbol{R}) = \frac{(Z_2 + Z_4)(Z_1 + Z_3)}{Z_1Z_2Z_3 + Z_1Z_2Z_4 + Z_1Z_3Z_4 + Z_2Z_3Z_4},$$

$$var(\hat{\beta}_2|\boldsymbol{R}) = \frac{(Z_3+Z_4)(Z_1+Z_2)}{Z_1Z_2Z_3+Z_1Z_2Z_4+Z_1Z_3Z_4+Z_2Z_3Z_4},$$

where $Z_k = \sum_{r\in G_k}(1-R_r)$ is the sum of the response indicators for Group G_k, $k=1,\ldots,4$. The A-optimal design for this model minimises an appropriate approximation to

$$E\{var(\hat{\beta}_0|\boldsymbol{R})\} + E\{var(\hat{\beta}_1|\boldsymbol{R})\} + E\{var(\hat{\beta}_2|\boldsymbol{R})\}$$

subject to the constraints $\sum_{k=1}^{4} w_k = 1$ (equivalent to the constraint $n_1+n_2+n_3+n_4=n$) and $w_k \geq 0,\ k=1,\ldots,4$. See Appendix A.4 for the analytical expression of the objective function for A-optimality. The corresponding expression for D-optimality is not given here, but it can be easily obtained through the use of analytical software such as *Maple 17* or *Mathematica*.

Setting $n=291$ and using the above estimated MAR mechanism, the optimal design is found by using the *Minimize* function in *Mathematica*, subject to the weight constraint. Table 5.5 shows the allocation scheme of a A- and a D-optimal design, denoted by ξ_A^* and ξ_D^* respectively. In the example considered here, due to the large sample size, we did not find any significant differences between the designs obtained through the first and second order approximations and so we have not distinguished between both designs here. In addition, the probability the regression coefficients cannot be estimated here is small and are less than 10^{-20}, so there is no significant drawback using the 1st order approximation.

Table 5.5: A- and D-optimal designs for the Alzheimer's example. The numbers in parentheses indicate the expected number of missing values in the respective group.

	n_1 w_1	n_2 w_2	n_3 w_3	n_4 w_4	n
ξ_A^*	108(61.1)	64(29.6)	64(22.2)	55(14.3)	291
	0.371	0.220	0.219	0.190	
ξ_D^*	60(33.9)	72(33.4)	78 (27.0)	81(21.1)	291
	0.206	0.248	0.268	0.278	

Using the same procedure as in Section 4, we assess the performance of the optimal designs by simulating incomplete data from the different designs using (5.1) above, choosing values of $\beta_0, \beta_1, \beta_2$ to be $1, 1, 1$ respectively. The missing values are introduced into the response using the MAR mechanism specified above with parameters $\hat{\gamma}_0 = 0.2636472$, $\hat{\gamma}_1 = -0.8988845$ and $\hat{\gamma}_2 = -0.4108504$ estimated from the data. From each incomplete data set, regression coefficients $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ are estimated from the complete cases. We repeat this process 350000 times to generate 350000 incomplete data sets, which allows us to empirically obtain the covariance matrix for $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ for each design across the replications. The original design, i.e. $\xi_{ori} = (n_1, n_2, n_3, n_4) = (72, 74, 73, 72)$ with expected missing observations $(40.7, 34.3, 25.3, 18.7)$ is also considered as a candidate design here.

Table 5.6 presents the simulated values for the A- and the D-objective function for the different designs. As expected, ξ_A^* has the smallest value for $tr\left(\text{sample } \boldsymbol{var(\hat{\beta})}\right)$ as this optimal design minimises the trace of our approximation to this matrix. Similarly, ξ_D^* has the smallest determinant of the simulated covariance matrix. Both designs result in an improved criterion value over the original design used and so could potentially have improved performance if they had been applied. For example, the A-optimal design would be expected to achieve a similar trace of the sample covariance matrix as the original design, while requiring only 95.55% of the overall sample size, or 13 fewer patients.

Table 5.6: Simulated values for the A- and the D-objective function, respectively, for different designs.

	$tr\left(\text{sample } \boldsymbol{var(\hat{\beta})}\right)$	$\lvert\text{sample } \boldsymbol{var(\hat{\beta})}\rvert$
ξ_A^*	**0.066327**	3.722e-06
ξ_D^*	0.072111	**3.3028e-06**
ξ_{ori}	0.069416	3.3439e-06

6. Discussion and remarks

In this article we have proposed a theoretical framework for designing experiments that takes into account the possibility of missing values. By incorporating a model for the missing data mechanism, we are able to create designs that optimise an objective function of interest. For various commonly used objective functions we have shown that our designs have the potential to improve performance over designs that do not necessarily incorporate the missing data model into the optimisation.

Our framework has broadened the approach proposed by Imhof, Song and Wong (2002), which is in fact a special case of the framework suggested here that only takes a Taylor expansion of order one. We have illustrated the potential benefits of extending the first order approach through a simulation study. In some situations, in particular for large sample sizes, the first and second order approach tend to lead to very similar designs. We can see how this occurs from the results in (3.5) and (3.6) as the terms arising from the second order expansion are of $O(n^{-2})$ and will decrease the fastest as the sample size increases. In these situations the first order approach might be preferred for practical reasons. The sample size of 30 we considered in Section 4 is typical for Phase II clinical trials, where sample sizes are normally no more than 50. In this situation our investigation in Section 4 showed that the 2nd order approximation offered various benefits over the 1st order. We have also noted some further theoretical properties of using an approach based on the first order expansion and derived the necessary results in this article.

We have described our methodology for the general linear regression model, and have illustrated its benefits through one- and two-variable models for simplicity. In these situations, the necessary Taylor expansions could easily be derived by hand. For more complicated linear models, in particular if the size of the covariance matrix is large, it is recommended to use symbolic computation software, such as Mathematica, for deriving the second order approximation to the covariance matrix. Numerical computation of optimal designs will be challenging since convexity of the objective function is not guaranteed, but is feasible e.g. using metaheuristic search algorithms such as PSO; see, e.g., Chen et al. (2015).

Our methodology is also applicable to nonlinear and generalised linear regression models. For nonlinear regression models with normally distributed errors, this can readily be seen by considering linearisation of the regression function; see e.g. Atkinson, Donev and Tobias (2007), Chapter 17.2. More generally, the formula $\boldsymbol{var}(\hat{\boldsymbol{\beta}}) \approx \boldsymbol{E}(\boldsymbol{var}(\hat{\boldsymbol{\beta}}|\boldsymbol{R}))$ is also applicable here where $\boldsymbol{var}(\hat{\boldsymbol{\beta}}|\boldsymbol{R})$ is replaced by the inverse of the Fisher information matrix conditional on the random indicators for missingness. We note that in these situations, the conditional covariance matrix is not available in closed form, but has already been approximated, which adds another level of approximation.

We note that accounting for the impact caused by the MAR mechanism in the optimal design framework does not entirely overcome the potential issues caused by singular information matrices. In situations where completely observed responses are not possible, Imhof, Song and Wong (2002) have shown that designs found by the first order approximation approach have some advantages over the conventional optimal designs. This has been confirmed by our investigations in Section 4. In addition, we have provided evidence that the second order approximation leads to optimal designs which perform better in this respect, i.e. the probability of a singular covariance matrix is smaller than for optimal designs using the first order approximation. In view of this, we suggest to compare the optimal designs found for the different approaches prior to the implementation of the experiment.

So far we have assumed that the analysis will be performed under a complete case analysis. While for many types of models such as linear regression models under a MAR mechanism, parameter estimates will be unbiased, this may not necessarily be the best way to handle the missing value problem. Another common approach to handling the missing data problem is by multiply imputing the missing values. Analysing the incomplete data in this way will not necessarily lead to the same designs derived in this article and further work will be needed to determine what the optimal designs will be here, as well as comparing which approach will lead to improved performance.

We note that the proposed framework here relies on the assumption of MAR, but it may be the case that missingness is not missing at random (NMAR). NMAR is a very challenging problem that poses additional challenges in constructing designs. For example, the dependency of the missing mechanism on the outcome means that the optimal design could be sensitive to model parameters that typically do not determine the optimal design when missingness is MAR. This is something that merits future investigation.

Acknowledgement

The first author's research is funded by the Institute for Life Sciences at the University of Southampton. We would like to acknowledge Clive Holmes; Robert Howard and Patrick Philips for supplying us with the data from the Domino study RCTN49545035 which was funded by the MRC and Alzheimer's Society UK.

Appendix

A.1 Proof of Theorem 1. We can prove that the D-optimal design has $p+1$ support points using the general equivalence theorem, by finding a contradiction. Assume ξ^* has $p+2$ support points. Consider

$$g(x) := \frac{\boldsymbol{f}^T(x)\ \boldsymbol{M}^{-1}(\xi^*)\ \boldsymbol{f}(x)}{p+1} \leq \frac{1}{1-P(x)} := h(x)$$

where $g(x)$ is a polynomial of degree $2p$, which has to be less than $h(x)$ over the region $[l,u]$. We order the $p+2$ values for x by size:

$$l \leq x_1^* < x_2^* < ... < x_{p+2}^* \leq u \tag{1}$$

such that the above equality is achieved. This implies $g(x_i^*)$ touches $h(x_i^*)$ and $g^{'}(x_i^*) = h^{'}(x_i^*)$ for $i = 2, 3, ..., x_{p+1}^*$. From (1), there are values $x_1^{*'}, ..., x_{p+1}^{*'}$ with $g^{'}(x_i^{*'}) = h^{'}(x_i^{*'})$ such that $x_1^* < x_1^{*'} < x_2^* < x_2^{*'} < x_3^* < ... < x_{p+1}^* < x_{p+1}^{*'} < x_{p+2}^*$ by the Mean Value Theorem.

Hence we have a total of $2p+1$ values where g and h have equal derivatives, and $g^{'}(x)$ is a polynomial of degree $2p-1$. Applying the Mean Value Theorem again to $g^{'}$ and $h^{'}$, there must be $2p$ values where $g^{''}$ and $h^{''}$ are equal. By repeating this process, we find that there must be 2 values where the $2p^{th}$ derivatives $g^{(2p)}$ and $h^{(2p)}$ are equal, and $g^{(2p)}(x)$ is a constant since g is a polynomial of degree $2p$.

This is a contradiction since we assumed that $h^{(2p)}(x) = c$ has at most one solution in $\Re$ for any constant c. The same contradiction occurs if we assume ξ^* has more than $p+2$ support points. □

A.2 Multivariate second order Taylor series approximation. Let X and Y be two independent discrete random variables with expectations $\overline{X}$ and $\overline{Y}$, respectively. Define $F = X$ and $G = XY$, which have expected values $\overline{X}$ and $\overline{XY}$ respectively. Assume we want to expand $H(F,G) = F/G$ about the point $(\overline{X}, \overline{XY})$ into a multivariate second order Taylor series. The partial derivatives evaluated at the point $(\overline{X}, \overline{XY})$ are

$$\frac{\partial\ H(\overline{X},\overline{XY})}{\partial F} = \frac{1}{\overline{XY}}\ ; \qquad \frac{\partial^2\ H(\overline{X},\overline{XY})}{\partial F^2} = 0; \qquad \frac{\partial^2\ H(\overline{X},\overline{XY})}{\partial F\ \partial G} = -\frac{1}{(\overline{XY})^2};$$

$$\frac{\partial\ H(\overline{X},\overline{XY})}{\partial G} = -\frac{\overline{X}}{(\overline{XY})^2}; \qquad \frac{\partial^2\ H(\overline{X},\overline{XY})}{\partial G^2} = 2\frac{\overline{X}}{(\overline{XY})^3}\ ; \qquad \frac{\partial^2\ H(\overline{X},\overline{XY})}{\partial G\ \partial F} = -\frac{1}{(\overline{XY})^2}.$$

Thus, a second-order Taylor series expansion for the function $H(F,G)$ expanded about the point $(\overline{X},\overline{XY})$ is

$$\begin{aligned}
&H(F,G)\\
&\approx H(\overline{X},\overline{XY}) + (F-\overline{X})\frac{\partial\ H(\overline{X},\overline{XY})}{\partial F} + (G-\overline{XY})\frac{\partial\ H(\overline{X},\overline{XY})}{\partial G}\\
&\quad + \frac{1}{2}\Bigg((F-\overline{X})^2 \underbrace{\frac{\partial^2\ H(\overline{X},\overline{XY})}{\partial F^2}}_{=0} + (G-\overline{XY})^2\frac{\partial^2\ H(\overline{X},\overline{XY})}{\partial G^2}\\
&\qquad\qquad +2(F-\overline{X})(G-\overline{XY})\frac{\partial^2\ H(\overline{X},\overline{XY})}{\partial G\ \partial F}\Bigg)\\
&= \frac{\overline{X}}{\overline{XY}} + (F-\overline{X})\left(\frac{1}{\overline{XY}}\right) - (G-\overline{XY})\left(\frac{\overline{X}}{(\overline{XY})^2}\right) + (G-\overline{XY})^2\left(\frac{\overline{X}}{(\overline{XY})^3}\right)\\
&\quad - (F-\overline{X})(G-\overline{XY})\left(\frac{1}{(\overline{XY})^2}\right).
\end{aligned}$$

To construct an optimal design, the expected value of the approximated function expanded about the point $(\overline{X},\overline{XY})$ is required. Since $E\{(F-\overline{X})\}=0$ and $E\{(G-\overline{XY})\}=0$,

$$\begin{aligned}
&E\{H(F,G)\}\\
&\approx \frac{\overline{X}}{\overline{XY}} + E\{(G-\overline{XY})^2\}\left(\frac{\overline{X}}{(\overline{XY})^3}\right) - E\{(F-\overline{X})(G-\overline{XY})\}\left(\frac{1}{(\overline{XY})^2}\right)\\
&= \frac{\overline{X}}{\overline{XY}} + (E\{G^2\}-(\overline{XY})^2)\frac{\overline{X}}{(\overline{XY})^3} - (E\{FG\}-\overline{X}\ \overline{XY})\frac{1}{(\overline{XY})^2}\\
&= E\{G^2\}\frac{\overline{X}}{(\overline{XY})^3} - \frac{E\{FG\}}{(\overline{XY})^2} + \frac{\overline{X}}{\overline{XY}} \;=\; \frac{E\{G^2\}E\{F\}}{E\{G\}^3} - \frac{E\{FG\}}{E\{G\}^2} + \frac{E\{F\}}{E\{G\}}.
\end{aligned}$$

Approximating the elements of $\boldsymbol{E}\{[\boldsymbol{M}(\xi,\boldsymbol{R})]^{-1}\}$ of the general linear model. Considering F and G are the products of independent binomial random variables present in the elements of $\boldsymbol{var(\hat{\beta})}$, the value of $\boldsymbol{E}\{[\boldsymbol{M}(\xi,\boldsymbol{R})]^{-1}\}$ can be approximated. For instance, the matrix for a simple linear model contains $E\{Z_i/(Z_iZ_j)\}$ where $i\neq j$; for a quadratic model, this matrix contains $E\{Z_iZ_j/(Z_iZ_jZ_k)\}$ where no pair of i,j,k is equal. Rewriting $G=FZ$ where Z

is the extra independent variable, we have

$$
\begin{aligned}
E\left\{H(F,G)\right\} &\approx \frac{E\{F^2\}E\{Z^2\}E\{F\}}{E\{F\}^3E\{Z\}^3} - \frac{E\{F^2\}E\{Z\}}{E\{F\}^2E\{Z\}^2} + \frac{E\{F\}}{E\{F\}E\{Z\}} \\
&= \frac{E\{F^2\}E\{F\}\left(E\{Z^2\} - E\{Z\}^2\right)}{E\{F\}^3E\{Z\}^3} + \frac{1}{E\{Z\}} \\
&= \frac{E\{F^2\}Var(Z)}{E\{F\}^2E\{Z\}^3} + \frac{1}{E\{Z\}}.
\end{aligned}
$$

Example: Approximation of $E\{Z_i/(Z_iZ_j)\}$ for $i,j = 1,2,\ i \neq j$. Z_i is a binomial random variable with mean $nw_i(1-P(x_i^*))$ and variance $nw_i(1-P(x_i^*))P(x_i^*)$, let $F = Z_i$ and $G = Z_iZ_j = FZ_j$. Using the above expression where the extra variable is Z_j in this example, we have

$$
\begin{aligned}
&E\left(\frac{Z_i}{Z_iZ_j}\right) \approx \frac{1}{E\{Z_j\}} + \frac{E\{Z_i^2\}Var(Z_j)}{(E\{Z_i\})^2(E\{Z_j\})^3} \\
&= \frac{1}{nw_j(1-P(x_j^*))} + \frac{(nw_i(1-P(x_i^*))P(x_i^*)+(nw_i(1-P(x_i^*)))^2)nw_j(1-P(x_j^*))P(x_j^*)}{(nw_i(1-P(x_i^*)))^2(nw_j(1-P(x_j^*)))^3} \\
&= \frac{1}{nw_j(1-P(x_j^*))} + \frac{P(x_i^*)P(x_j^*)}{nw_i(1-P(x_i^*))(nw_j(1-P(x_j^*)))^2} + \frac{P(x_j^*)}{(nw_j(1-P(x_j^*)))^2}
\end{aligned}
$$

A.3 Proof of part (a) of Theorem 3. We substitute $w_1 = 1 - w_2$ into the respective objective function and take the partial derivative with respect to w_2. For the D-objective function, we obtain

$$
\begin{aligned}
&\tfrac{\partial}{\partial w_2}\left(E\left(\tfrac{Z_1}{Z_1Z_2}\right)E\left(\tfrac{Z_2}{Z_1Z_2}\right)\right) \\
&= \left(\tfrac{\partial}{\partial w_2}E\left(\tfrac{Z_1}{Z_1Z_2}\right)\right)E\left(\tfrac{Z_2}{Z_1Z_2}\right) + E\left(\tfrac{Z_1}{Z_1Z_2}\right)\left(\tfrac{\partial}{\partial w_2}E\left(\tfrac{Z_2}{Z_1Z_2}\right)\right) \\
&\approx \frac{(2\,w_2-1)\left(n'^4{w_2}^4-2\,n'^4{w_2}^3-6\,P^2n'^2{w_2}^2-2\,Pn'^3{w_2}^2+n'^4{w_2}^2+6\,P^2n'^2w_2+2\,Pn'^3w_2+3\,P^4+3\,P^3n'\right)}{n'^6(-1+w_2)^4{w_2}^4}
\end{aligned}
$$

and the optimal weight $w_2 = 1/2$ as $\tfrac{\partial}{\partial w_2}\left(E\left(\tfrac{Z_1}{Z_1Z_2}\right)E\left(\tfrac{Z_2}{Z_1Z_2}\right)\right) = 0$. On the other hand, the derivative of the c-objective function with respect to w_2 is

$$\begin{aligned}&\frac{\partial}{\partial w_2}E\left(\frac{Z_1}{Z_1Z_2}\right)+\frac{\partial}{\partial w_2}E\left(\frac{Z_2}{Z_1Z_2}\right)\\&\approx -\frac{\left(2\,w_2-1\right)\left(2\,Pn'{w_2}^2-{n'}^2{w_2}^2-2\,Pn'w_2+{n'}^2w_2+2\,P^2+2\,Pn'\right)}{{n'}^3{w_2}^3\left(-1+w_2\right)^3},\end{aligned}$$

which yields that the optimal weight is $w_2 = 1/2$ for this criterion as well. Hence, we have shown that the optimal weight for the experiment with MCAR mechanism is the same as the optimal weight for complete observations.

From (3.6), for both optimality criteria, the expressions above do not depend on the support points. Hence the objective functions in (3.7) and (3.8), respectively, are minimised with respect to x_1^* and x_2^* when the factor $1/(x_1^* - x_2^*)^2$ is minimised. This is achieved by setting $x_1^* = l$ and $x_2^* = u$.

Taking partial derivatives in (3.9) with respect to x_1^* and x_2^*, respectively, shows that regardless of the values of the expression in (3.6) the derivative with respect to x_1^* is non-negative if $l \geq 0$ or $u \leq 0$. Hence the A-objective function is minimised when $x_1^* = l$. Similarly, the derivative with respect to x_2^* is non positive if $l \geq 0$ or $u \leq 0$. Hence the A-objective function is minimised when $x_2^* = u$. □

A.4 The covariance matrix of model $Y_i = \beta_0 + \beta_1 x_i + \beta_2 v_i + N(0, \sigma^2)$ from the Alzheimer's example

$$[\boldsymbol{M}(\xi,\boldsymbol{R})]^{-1}=\frac{1}{|\boldsymbol{M}(\xi,\boldsymbol{R})|}\begin{pmatrix} Z_2Z_3+Z_2Z_4+Z_4Z_3 & -(Z_2+Z_4)Z_3 & -(Z_3+Z_4)Z_2\\ -(Z_2+Z_4)Z_3 & (Z_2+Z_4)(Z_1+Z_3) & -Z_4Z_1-Z_2Z_3\\ -(Z_3+Z_4)Z_2 & -Z_4Z_1-Z_2Z_3 & (Z_3+Z_4)(Z_1+Z_2)\end{pmatrix}$$

where $|\boldsymbol{M}(\xi,\boldsymbol{R})| = Z_1Z_2Z_3 + Z_1Z_2Z_4 + Z_1Z_3Z_4 + Z_2Z_3Z_4$. After some simplification, the trace is found to be

$$\frac{Z_1Z_2+Z_1Z_3+2\,Z_1Z_4+3\,Z_2Z_3+2\,Z_2Z_4+2\,Z_3Z_4}{Z_1Z_2Z_3+Z_1Z_2Z_4+Z_1Z_3Z_4+Z_2Z_3Z_4}$$

where $Z_k = \sum_{i\in G_k}(1 - R_i)$ is the sum of the response indicators in Group $G_k, k = 1, \ldots, 4$. Using Taylor second order linearisation where

$$E\left(\frac{F}{G}\right)\approx\frac{E\{G^2\}E\{F\}}{(E\{G\})^3}-\frac{E\{FG\}}{(E\{G\})^2}+\frac{E\{F\}}{E\{G\}},$$

the A-objective function can now be found by taking the expectation with respect to the binomial random variables where $E(Z_k) = nw_k(1 - P(x_k^*))$ and $E(Z_k^2) = nw_k(1 - P(x_k^*))P(x_k^*) + (E(Z_k))^2$ for $k = 1, 2, 3, 4$. The expressions F and G are given by $F = Z_1Z_2 + Z_1Z_3 + 2\,Z_1Z_4 + 3\,Z_2Z_3 + 2\,Z_2Z_4 + 2\,Z_3Z_4$ and $G = Z_1Z_2Z_3 + Z_1Z_2Z_4 + Z_1Z_3Z_4 + Z_2Z_3Z_4$.

References

Ahmad, T., and Gilmour, S. G. (2010). Robustness of subset response surface designs to missing observations. *Journal of Statistical Planning and Inference*, **140(1)**, 92-103.

Baek, I., Zhu, W., Wu, X., and Wong, W. K. (2006). Bayesian optimal designs for a quantal dose-response study with potentially missing observations. *Journal of Biopharmaceutical Statistics*, **16(5)**, 679-693.

Bang, H., and Robins, M. J. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, **61(4)**, 962-973.

Carpenter, J. R., Kenward, M. G., and White, I. R. (2007). Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical Methods in Medical Research*, **16(3)**, 259-275.

Chen, R. B., Chang, S. P., Wang, W., Tung, H. C. and Wong, W. K. (2015). Minimax optimal designs via particle swarm optimization methods. *Statistics and Computing*, **25(5)**, 975-988.

De la Garza, A. (1954). Spacing of information in polynomial regression. *The Annals of Mathematical Statistics*, **25(1)**, 123-130.

Fedorov, V. V. (1972). Theory of optimal experiments. *Elsevier*

Ghosh, S. (1979). On robustness of designs against incomplete data. Sankhyā: *The Indian Journal of Statistics, Series B*, 204-208.

Glynn, R. J., and N. M. Laird. (1986). Regression estimates and missing data: complete case analysis. *Technical Report: Harvard School of Public Health, Department of Biostatistics.*

Hackl, P. (1995). Optimal design for experiments with potentially failing trials. In *Proc. of MODA4: Advances in Model-Oriented Data Analysis* (Edited by C. P. Kitsos and W. G. Müller), 117-124. Physica Verlag, Heidelberg.

Hedayat, A. and John, P. W. M. (1974). Resistant and susceptible BIB designs. *The Annals of Statistics* , **2(1)**, 148–158.

Herzberg, A. M. and Andrews, D. F. (1976). Some considerations in the optimal design of experiments in non-optimal situations. *Journal of the Royal Statistical Society: Series B (Methodological)*, **38**, 284-289.

Howard, R., McShane, R., Lindesay, J., Ritchie, C., Baldwin, A., Barber, R., ... and Phillips, P. (2012). Donepezil and memantine for moderate-to-severe Alzheimer's disease. *New England Journal of Medicine*, **366(10)**, 893-903.

Ibrahim, J. G. and Lipsitz, S. R. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society: Series B (Methodological)*, **61(1)**, 173-190.

Imhof, L. A and Song, D. and Wong, W. K. (2002). Optimal design of experiments with possibly failing trials. *Statistica Sinica*, 1145-1155.

Imhof, L. A and Song, D. and Wong, W. K. (2004). Optimal design of experiments with anticipated pattern of missing observations. *Journal of Theoretical Biology*, **228(2)**, 251-260.

Kenward, M. G., Molenberghs, G., and Thijs, H. (2003). Pattern mixture models with proper time dependence. *Biometrika*, **90(1)**, 53-71.

Little, R. J. A. (1992). Regression with missing X's: a review. *Journal of the American Statistical Association*, **87(420)**, 1227-1237.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, **90(431)**, 1112-1121.

Little, R. J. A. and Rubin, D. B. (2002). Statistical analysis with missing data. *J. Wiley.*

Mitra, R. and Reiter, J.P. (2011). Estimating propensity scores with missing covariate data using general location mixture models. *Statistics in Medicine*, **30(6)**, 627-641.

Mitra, R. and Reiter, J.P. (2016). A comparison of two methods of estimating propensity scores after multiple imputation. *Statistical Methods in Medical Research*, **25(1)**, 188-204.

Ortega-Azurduy, S. A., Tan, F. E. S. and Berger, M. P. F. (2008). The effect of dropout on the efficiency of D-optimal designs of linear mixed models. *Statistics in Medicine*, **27(14)**, 2601-2617.

Pukelsheim, F. (2006). Optimal design of experiments. *Society for Industrial and Applied Mathematics.*

Pukelsheim, F. and Rieder, S. (1992). Efficient rounding of approximate designs. *Biometrika*, **79(4)**, 763-770.

Rubin, D.B (1976). Inference and missing data. *Biometrika*, **63(3), 581-592.**

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. *Wiley-Interscience.*

Schafer, J. L. (1997). Analysis of incomplete multivariate data. *CRC press.*

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

Spratt, M., Carpenter, J., Sterne, J. A., Carlin, J. B., Heron, J., Henderson, J., and Tilling, K. (2010). Strategies for multiple imputation in longitudinal studies. *American Journal of Epidemiology*, **172(4)**, 478-487.

White, I. R., Higgins, J., and Wood, A. M. (2008). Allowing for uncertainty due to missing data in meta-analysis—Part 1: Two-stage methods. *Statistics in Medicine*, **27(5)**, 711-727.