1. Introduction

Information Capacity (IC) is a criterion for selecting a design for an experiment that maximises the average estimation efficiency across a specified set of models, \( M \), to be investigated in the data analysis (see Li and Nachtsheim [1]). This paper describes how this criterion can be applied to an experiment of \( N \) runs where a generalised linear model (GLM) describes the measured response. Three different types of designs are compared with the aim of achieving accurate estimation of the models in the set, and also the ability to discriminate between these competing models.

2. Methods

Three types of design structure are investigated:

- **Unrestricted designs** - designs where the factor values can be any point in the design region \( \mathcal{X} \);
- **Restricted designs** - designs with each factor restricted to three pre-specified values: -1, 0, and 1; and
- **Fractional factorial designs**.

Designs are assessed through:

- Examination of their efficiency for estimating the parameters in each of the models in set \( M \), and
- Simulation studies to examine their ability to discriminate between the models in \( M \).

3. Background

3.1 Generalized Linear Models

In a GLM, observations are independent and described by a response distribution from the exponential family, a linear predictor, \( \mathcal{X} \) (where \( \beta \) is the vector of unknown parameters and \( X \) is the model matrix), and a link function (see McCullagh and Nelder [2]).

3.2 Designs

A design \( \zeta \) defines the \( n \) distinct combinations \( \{ x_i \in \mathcal{X} \} \) of values of the \( j \) factors, called support points, to be run in the experiment, together with the proportion \( w_i \) of the total experimental effort expended at the \( j \)th point \( \{ 1, \ldots, n \}; \quad 0 < w_i \leq 1, \sum_i w_i = 1 \).

For an exact design, each \( w_i \) must be integer; otherwise \( \zeta \) is called an approximate design.

4. Information Capacity Designs

To illustrate the efficiency of IC-optimal designs, we consider logistic regression with a first-order linear predictor. Five factors may possibly affect the response and \( M \) consists of \( 2^5 - 1 \) possible models (excluding the null model). The parameter space for a model is defined from

\[
0.5 \leq \beta_j \leq 1.5, \quad j = 1, \ldots, 5
\]

where the first column specifies that the intercept is 0, and the \( (j+1) \)th column defines the range of values of the coefficient of the \( j \)th variable \( (j = 1, \ldots, 5) \); each range includes zero.

We used a numerical approximation to the integral in equation (1) and computer search to find two exact IC-optimal designs in \( N = 16 \) runs: an unrestricted, and a restricted design (offering computational advantage). For these designs and a fractional factorial (FF) design, the median and minimum D-efficiencies were calculated for each \( \beta \in \mathcal{M} \) using 500 random draws from \( \beta_0 \). The fraction has defining contrast \( \zeta = (1, 2, 1, 1, 1, 1, 1, 1) \), where \( A, \ldots, I \) are the factor labels. Figure 1 shows that the IC-optimal designs performed consistently better than the FF design.

5. Model Estimation

6. Model Discrimination

To assess the effectiveness of the IC-optimal designs and evaluate their performance for choosing between models, a simulation study was conducted, where model selection was via AIC. A simulation run consisted of data generation from each of the \( 2^5 \) possible models and model selection. 1000 simulation runs were performed. Four different performance measures were recorded:

- **Coverage** (proportion of simulation runs where the designated 'best' model is correctly identified);
- **Power** (the average proportion of active factors correctly identified);
- **Type 1 error rate** (the average proportion of inactive factors incorrectly declared active); and
- **False discovery rate** (the average proportion of declared active factors which are actually inactive).

6.1 Exact Designs

The results of the simulation study are summarized in Figure 2 by averaging the performance measures over the subset of models having \( j \) factors \( (j = 1, \ldots, 5) \). It can be seen that the FF design has the best performance for all measures except type 1 error rate and false discovery rate for models with exactly 2 factors. The unrestricted design is better than the restricted design for type 1 error rate and false discovery rate. These two types of designs have similar performance for power and coverage.

7. Discussion

The findings from our studies indicate that:

- The IC-optimality criterion can be used to find efficient designs for flexible numbers of factors and runs;
- The unrestricted and restricted IC-optimal designs have similar estimation efficiency for different ranges of parameters across a set of models;
- IC-optimal designs have better performance than a fractional factorial design for parameter estimation;
- IC-optimal designs are effective for model discrimination particularly for larger numbers of runs.

References