Statistical Methodology 32 (2016) 36–52

Contents lists available at ScienceDirect

Statistical Methodology

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

Bayesian optimal cluster designs

Satya Prakash Singh, Siuli Mukhopadhyay *

Department of Mathematics, Indian Institute of Technology Bombay, Mumbai, India

A R T I C L E I N F O

Article history:
Received 21 November 2014
Received in revised form
25 January 2016
Accepted 22 February 2016
Available online 11 March 2016

Keywords:
Bayesian designs
Cost function
Intracluster correlation
Multi-objective optimization
Pareto optimality

A B S T R A C T

Designing cluster trials depends on the knowledge of the intracluster correlation coefficient. To overcome the issue of parameter dependence, Bayesian designs are proposed for two level models with and without covariates. These designs minimize the variance of the treatment contrast under certain cost constraints. A pseudo Bayesian design approach is advocated that integrates and averages the objective function over a prior distribution of the intracluster correlation coefficient. Theoretical results on the Bayesian criterion are noted when the intracluster correlation follows a uniform distribution. Two data sets based on educational surveys conducted in schools are used to illustrate the proposed methodology.

1. Introduction

Cluster or nested trials are widely popular in the health research community, educational surveys and social studies, though they require higher sample sizes to obtain similar power as compared to completely randomized studies [14]. The similarity of individuals in a cluster makes them less efficient, in terms of power and sample size. However from ethical, costs and contamination points [14], these designs are still extensively used. Designing cluster trials depends on the knowledge of the intracluster correlation coefficient. This poses a challenge to the researchers since at the planning stage nothing is known about the correlation. Several authors, like Raudenbush [33], Moerbeek et al. [26], Roy et al. [34], Heo and Leon [20] and Konstantopoulos [23], have proposed various design selection methods for cluster trials. However, all these methods are concerned with determining designs which are based on some initial estimates of the unknown parameters (correlation or variance components) or assuming known values based on prior studies.

* Corresponding author. Tel.: +91 2225767495.
E-mail addresses: spsingh@math.iitb.ac.in (S.P. Singh), siuli@math.iitb.ac.in (S. Mukhopadhyay).

http://dx.doi.org/10.1016/j.stamet.2016.02.002
1572-3127/

To overcome the issue of the parameter dependence in a design selection problem, Bayesian and minimax designs have been used frequently in the literature [3]. In a Bayesian design a prior distribution is assumed for the unknown parameters, which is then incorporated into an appropriate objective function by integrating and averaging over the prior distribution. For example, in Bayesian designs one criterion maximizes the average over the prior distribution of the logarithm of the determinant of Fisher's information matrix, making it equivalent to D-optimality in linear models. Bayesian versions of other alphabetic optimality criteria are also available in the literature. An early paper on Bayesian D-optimality criterion is the one by Zacks [40]. Later, the Bayesian approach has been discussed by several authors including [7–9] in the context of nonlinear and generalized linear models. Designs for exponential growth models, block designs and split plot designs were presented by Dette and Sperlich [13], Woods and Van de Ven [39] and Mylona et al. [31], respectively. D- and Ds-(optimal for a subset of parameters) optimal Bayesian designs for a compartmental mode can be found in [2]. Minimax/maximin designs are found by minimizing (maximizing) the maximum (minimum) of an objective function, for example the variance function, over a range of parameter values (see e.g., [32, 35,28,29]). The resultant designs, are particularly attractive since the experimenter is only required to specify an appropriate range for the unknown parameters [12]. Recently, the maximin/minimax approach to select robust optimal cluster designs was proposed by Van Breukelen and Candel [38].

In this article we propose Bayesian cluster designs by assuming a prior distribution on the intracluster correlation coefficient, which is then incorporated into the variance of the treatment contrast by integrating and averaging over the assumed prior. The performance of the Bayesian cluster designs proposed is also compared with the minimax designs of Van Breukelen and Candel [38] through several numerical examples.

To obtain a sufficient power value of the test for correctly judging the hypothesis a large number of subjects as well as clusters are required leading to an increase in the costs. Hence, there is a need to design cluster studies with adequate power while keeping the budget constraints in mind. Various methods for selecting cluster designs with adequate power under cost constraints have been proposed by Raudenbush [33], Moerbeek et al. [26], Roy et al. [34], Konstantopoulos [23]. Several books discussing selection of cluster designs [5,25] are also available. In this article, we look at Bayesian designs under both fixed and varying cost constraints. For varying cost, we use the concept of Pareto optimality in design selection.

In Section 2, two level without and with covariates models are defined and expressions for the power functions and variances of treatment contrast are given. The designs considered are balanced in the number of clusters and individuals. In Section 3, we define the locally optimal cluster design and provide a motivating example for Bayesian designs. The Bayesian criterion and a short discussion on the priors used for the intracluster correlation coefficient are given in Section 4. In Section 5, two numerical examples are given. Section 6 is concerned with selecting Bayesian designs under varying cost constraints. The relevancy of our proposed methodology and directions for future research are given in Section 7.

2. A two level nested model

We consider nested/clustered designs involving two factors one of which is nested within the other factor. Suppose we denote the two factors as A and B, and assume that B is nested within A. Factor A is assumed to be a fixed effect while factor B is random. We may consider an example where we are interested in comparing the math scores of children from different schools, when two different teaching methods are employed. The teaching method represents factor A and it has two levels, while the schools nested under the different teaching methods represent factor B. The model is [27],

$$y_{ijk} = \mu + \alpha_i + \beta_{ij} + \epsilon_{ijk} \tag{1}$$

where y_{ijk} is the math score of the kth child, $k = 1, 2, \ldots, n$, from the jth school, $j = 1, 2, \ldots, m$, which is nested under the ith teaching method. There are n subjects per cluster nested in m clusters in each group ($i = 1, 2$). The overall mean is μ, α_i is the fixed effect of the teaching method ($i = 1, 2$), β_{ij} is the random effect of the jth school nested within the ith teaching method, and ϵ_{ijk} is the random error. It is assumed that β_{ij} is independently distributed as $N(0, \sigma_\beta^2)$ and ϵ_{ijk} independently follows