Optimal Allocation with Known Covariates into Two Treatments under Generalized Linear Model

Technical Report No. ASU/2020/01

Dated 06 February, 2020

Samrat Hore, Department of Statistics, Tripura University

Anup Dewanji, Applied Statistics Unit, Indian Statistical Institute, Kolkata

Aditya Chatterjee, Department of Statistics, University of Calcutta

Optimal Allocation with Known Covariates into Two Treatments under Generalized Linear Model

Samrat Hore, Department of Statistics, Tripura University

Anup Dewanji, Applied Statistics Unit, Indian Statistical Institute, Kolkata

Aditya Chatterjee, Department of Statistics, University of Calcutta

Abstract

The problem of optimal allocation of experimental units with known covariates into several treatment groups under linear ANCOVA model has been discussed in several studies. Such design issue has been already discussed by the authors and the optimal allocation design has been derived through an efficient algorithm named as Hybrid variable neighborhood search (VNS) algorithm. In this paper, we have addressed the same issue with regard to $D-$ and $A-$optimality under the generalized linear model (GLM) set-up, assuming the response variable to be count, ordinal or binary, and the performance of the optimal allocation design obtained through the Hybrid VNS algorithm has been compared with respect to the random allocation scheme, under various GLM frameworks, through simulation studies and real-life examples.

Keywords: $D-$optimality, $A-$optimality, Hybrid Optimization, Mahalanobis Imbalance Metric, Constrained Optimality.

1. Introduction

In any allocation problem involving allocation of experimental units into two or more treatment groups, balancing of the covariates associated with the experimental units seems to be very important in all fields of scientific research, viz., clinical trials, agricultural experiments, chemical industry, medical research (Hu and Hu, 2012a, 2012b; Harville, 1974; Hinkelmann and Kempthorne, 2005,2007; Saville and Wood, 1991; Cartwright et al., 1968; Kalbfleish and Prentice, 2002, Rubin, 2008; Morgan and Rubin, 2012, to quote a few). Usually the linear model (LM) in ANCOVA set-up is considered for such problems and various optimality criteria have been addressed by several authors (Harville, 1974; Shah and Sinha, 1989, p 125-128; Hore et al., 2014, 2016). In theses works, close coincidence of optimality and balancing have been shown and demonstrated. However, in contrast with the usually assumed Gaussian response variable in the linear ANCOVA model, in many practical applications, the response variable is discrete (count, binary or categorical). In such cases, corresponding models are to be framed as the generalized linear model (GLM) (McCullagh and Nelder, 1989). In the present paper, we have tried to address the issue of optimum allocation of treatments with known covariates under such GLM set-up, vis-a-vis the issue of balancing.

Allocation problem related to GLM is more complex in comparison to the corresponding LM formulation, because the allocation design is dependent on the parameters of the model, which are unknown to us. To circumvent this problem, a common approach is to construct an efficient design for the best guess of the parameter values. Such approach is known as 'locally optimal' designs, as introduced by Chernoff (1953), and has been successfully implemented by Yang et al. (2012) and Yang and Mandal (2015) for $2-$level factorial design with binary response model and $D-$optimal factorial designs under the GLM frameworks, respectively. In quantal assay (Morgan, 1992) with binary response, there are several non-Bayesian approaches for efficient estimation of the model parameters which are known as initial estimate approach, sequential approach, constant information approach and fiducial approach (Finney, 1978; Abdelbasit and Plackett, 1983). At the same time, for such model, the Bayesian approach is discussed in Chaloner and Larntz (1989) and several other references cited therein. However, in both the approaches the basic objective is to find the optimum design points with regard to some optimality criteria. Wu (1985) has considered a Bayesian approach to obtain an optimal sequential allocation design for binary responses. Recently, Yang et al. (2016) suggested an innovative approach for achieving a $D-$optimal allocation design with the corresponding information matrix being replaced by its expectation with respect to prior distribution of parameters. This approach is known as the EW $D-$optimality criterion, where E stands for expectation and W for the appropriate weight function of the parameters for the usual $D-$optimality under various GLM frameworks. It is to be noted that this approach is similar to the Bayesian approach where the expectations of the key quantities, which are functions of the parameters, are evaluated with respect to some prior distributions of the parameters. In this paper, we have adopted this idea of $EW-$optimality for obtaining an efficient $D-$ and $A-$optimal allocation design. A kind of model-robust optimality has also been suggested in the following manner. The corresponding $D-$ and

$A-$optimal allocation designs are obtained by minimizing the maximum of the determinant and trace, respectively, of the dispersion matrix in which the maximum is taken over a set of model parameters randomly chosen from the assumed prior distribution and minimum is taken over all possible allocation designs. So, in a sense, it is a minimax design with an attempt to make it robust against mis-specification of values of the model parameters.

According to Rubin (2008), balanced allocation of experimental units with regard to various known covariates among several treatment groups, before the physical experiment takes place, is often considered to be the most reasonable allocation scheme in all intervention studies and clinical trials. Under the LM formulation, Hore et al. (2017) have shown that covariate mean balanced allocation is the necessary condition for achieving $D-$, $A-$, D_s- and A_s-optimality for known categorical covariates with multiple levels. Wherever, with known numerical covariates, Shah and Sinha (1989, p 125-128) and Antognini and Giovagnoli (2015, Appendix A) have shown that covariate mean balance across the treatment groups is a sufficient condition for attaining the $D-$, $A-$ and D_s-, A_s-optimality, respectively. This, however, need not be necessary as the covariate mean balance may not be achievable in many practical situations, especially for continuous covariates. As such, in place of nearly impossible and seldom achievable exact balance, minimizing an imbalance measure computed on the basis of known numerical covariates may be a way out to achieve the nearly balanced allocation design. To obtain a balanced covariate allocation, Morgan and Rubin (2012) suggested the re-randomization technique subject to achieving a minimum imbalance threshold with regard to the Mahalanobis imbalance metric (Imbens and Rubin, 2015, p 342). It has been noted by extensive simulations (Hore et al., 2014, 2016) that an optimal or near-optimal allocation design for known numerical covariates with regard to $D-$ and $A-$optimality may ensure covariate mean balance to the best possible extent under linear ANCOVA model. This finding in the LM set-up motivates us to explore the nature of balancing in similar situation under the GLM set-up, where the response variable is assumed to be binary or count. Interestingly, in this set-up, covariate mean balance is not ensured through such optimal allocation, as evident through our extensive simulation study (See Section 5). Similar observation was made in Rosenberger and Sverdlov (2008). According to them, for linear models with constant variance, the concepts of optimality and balance are frequently equivalent, but for nonlinear and heteroscedastic models they are not. Consequently, our next objective is to develop a robust allocation design that can retain a compromise between optimality and balancing, through some constrained optimization method.

The article has been structured in several sections. In Section 2, the problem has been formulated in the GLM framework and corresponding $D-$ and $A-$optimality in such situations are discussed. In the next section, optimality under Poisson GLM set-up for different prior distributions of the parameters are considered for a small number of experimental units and the performance of the allocation design obtained through the Hybrid VNS algorithm (Hore et al., 2014, 2016, 2018) is compared with respect to the exact optimal allocation obtained through complete search method. Efficiency under the same set-up, for a large number of experimental units has been investigated and reported in comparison to the randomized allocation rule. In Section 4, for binary response data with large number of units, we have compared the efficiency of the allocation design obtained by the proposed algorithm over the randomized allocation rule for various optimality criteria after generating samples from the prior distribution of the parameters through extensive simulation studies. In Section 5, the nature of imbalance with respect to the Mahalanobis metric matching (Imbens and Rubin, 2015, p 342) has been studied through simulation experiments while achieving the $D-$ and $A-$optimality under the usual LM and GLM frameworks with known covariates. It has been found that, with the same covariates, the balance with regard to the value of the Mahalanobis imbalance metric being nearly close to zero, with appropriate p-value very close to unity, is attained under the LM set-up but is not achieved in all kinds of GLM formulations. In this context, a compromise between balance and optimality has been studied in Section 6 to obtain a near-optimal allocation design, while ensuring a certain amount of balance through attaining a pre-fixed upper limit of the Mahalanobis imbalance metric. A real life example dealing with the Poisson regression model is discussed in Section 7. The paper ends up in Section 8 with some concluding remarks and scopes of further research.

2. Modeling and Optimality

Let us consider allocation of n experimental units with known $p-$dimensional covariates into two treatment groups. Suppose the $n-$ dimensional response vector Y has the mean ξ that depends on the linear predictor η, a linear function of the design matrix $\mathcal{U}$ and the vector of parameters, say θ. The relation between the response vector Y and the linear predictor η is modeled through the link function $g(\cdot)$ as (McCullagh and Nelder, 1989)

$$E(Y) = \xi \tag{2.1}$$

$$\eta = \mathcal{U}\theta \tag{2.2}$$

$$\eta_i = g(\xi_i), \tag{2.3}$$

where η_i and ξ_i are the i^{th} element of η and ξ, respectively, for $i = 1, \cdots, n$. In the context of two treatment groups and p covariates, the parameter vector $\theta = (\mu_1, \mu_2, \beta_1, \beta_2, \cdots, \beta_p)$ is a $(p+2)$ vector, with μ_1 and μ_2 being the two treatment effects and $\beta_1, \cdots, \beta_p$ being the p covariate effects, respectively. The design matrix $\mathcal{U}$ is of order $n \times (p+2)$, with the first two columns comprising of $1,$ or 0, depending upon the presence or absence of either of the the treatments, while the remaining p columns correspond to the values of the p covariates. To estimate the parameter θ, it is necessary that $n \geq (p+2)$. The corresponding dispersion matrix of the estimated parameter $\hat{\theta}$ is

$$Var(\hat{\theta}) = (\mathcal{U}^T W(\theta)\mathcal{U})^{-1}, \tag{2.4}$$

where $W(\theta) = Diag(w_1(\theta), w_2(\theta), \cdots, w_n(\theta))$ is a $n \times n$ diagonal matrix with the weight $w_i(\theta)$ being a function of $\eta_i = u_i^T\theta$, u_i being the i^{th} row of the design matrix $\mathcal{U}$, $i = 1, \cdots, n$. The functional form of the various link functions and the corresponding weights under several probability models are summarized in the following table (MacKenzie and Peng, 2014).

Table 1. Canonical link functions and corresponding weights.

Distribution	Density (mass) function	Link function ($g(\xi)$)	Weight ($w_i(\theta)$)
Normal	$f(y;\xi,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y-\xi)^2}{2\sigma^2}}$	ξ	σ^{-2}
Exponential	$f(y;\lambda) = \lambda e^{-\lambda y}$	ξ^{-1}	$(u_i\theta)^{-2}$
Poisson	$f(y;\lambda) = e^{-\lambda}\frac{\lambda^y}{y!}$	$log(\xi)$	$exp(u_i\theta)$
Bernoulli	$f(y;p) = p^y(1-p)^{1-y}$	$log(\frac{\xi}{1-\xi})$	$\frac{exp(u_i\theta)}{(1+exp(u_i\theta))^2}$
Binomial	$f(y;n_i,p) = \binom{n_i}{y}p^y(1-p)^{1-y}$	$log(\frac{\xi}{1-\xi})$	$\frac{n_i\ exp(u_i\theta)}{(1+exp(u_i\theta))^2}$

The $D-$ and $A-$optimal allocation designs for the efficient estimation of both treatment and covariate effects are to be obtained by minimization of the determinant and the trace of the dispersion matrix $(\mathcal{U}^T W(\theta)\mathcal{U})^{-1}$ (Kiefer, 1953). Note that this matrix, unlike the LM formulation, depends upon the parameter θ. To overcome this difficulty, the most convenient approach is to construct an optimal or near-optimal allocation design for the best guess of the parameter values, known as 'local optimality approach', introduced by Chernoff (1953) and implemented on $2-$level factorial design with binary response model and

$D-$optimal factorial design under the GLM frameworks by Yang et al. (2012) and Yang and Mandal (2015), respectively. An alternate to this 'local optimality approach', and named as *EW* $D-$optimality, has been suggested in Yang et al. (2016) to achieve $D-$optimal allocation design by minimizing $E_\theta Var(\hat{\theta})$ as the objective function, where the expectation is taken over the prior distribution of the parameter vector θ. We have adopted this idea of $EW-$optimality for obtaining an efficient $D-$ and $A-$optimal allocation design through the proposed Hybrid VNS algorithm. In practice, instead of minimization of the determinant or trace of the dispersion matrix $(\mathcal{U}^T W(\theta)\mathcal{U})^{-1}$ for obtaining the $D-$ and $A-$optimality, respectively, the expected value of the variance is computed prior to the minimization. These are called *EW* $D-$optimal and *EW* $A-$optimal allocation design, respectively. To evaluate the expectation of dispersion matrix, we require to assume a prior distribution for the parameter vector θ, say $\pi(\theta)$. The expectation $E_\theta Var(\hat{\theta}) = \int_\theta Var(\hat{\theta})\pi(\theta)d\theta$ may be evaluated exactly or through simulation, depending upon the link functions.

It is to be noted that, for $D-$optimality, we may consider minimization of either $D^{(1)} = |E_\theta Var(\hat{\theta})|$ or $D^{(2)} = E_\theta|Var(\hat{\theta})|$, with the obvious restriction of $D^{(1)} \leq D^{(2)}$, by Jensen's inequality (Yang et al., 2016, p 392). Similarly, for $A-$ optimality, it is possible to consider minimization of either $A^{(1)} = trace\ E_\theta Var(\hat{\theta})$ or $A^{(2)} = E_\theta\ trace\ Var(\hat{\theta})$, where $A^{(1)} = A^{(2)}$. In case the exact expression of the integral $\int_\theta Var(\hat{\theta})\pi(\theta)d\theta$ giving $E_\theta Var(\hat{\theta})$ is not available, we can approximate it through Monte-carlo (MC) approach, which invokes the law of large numbers for evaluation of complex integrals indicating the expected value of a functional under $\pi(\theta)$ through the computation of sample mean (i.e. empirical mean) of the functional on the basis of independent samples drawn from various $\pi(\theta)$ (Hastings, 1970). Thus $\int_\theta Var(\hat{\theta})\pi(\theta)d\theta$ is approximated by $\frac{1}{N}\sum\limits_{i=1}^{N} Var(\hat{\theta}_i)$, where $\hat{\theta}_i$, for $i = 1, 2, \cdots, N$, are N independent samples from $\pi(\theta)$ (Tanner, 1993). In this context, by following the robust allocation design concept of Hore et al. (2014), it is tempting to suggest one more pair of objective functions given by $D^{(3)} = max\{|Var(\hat{\theta}_1)|, |Var(\hat{\theta}_2)|, ..., |Var(\hat{\theta}_N)|\}$ and $A^{(3)} = max\{trace\ Var(\hat{\theta}_1), ..., trace\ Var(\hat{\theta}_N)\}$ which are to be minimized, to obtain something similar to $D-$ and $A-$optimality, respectively.

Suppose n number of experimental units with known $p-$covariates are to be allocated into two treatment groups under Poisson GLM set-up, where n_l number of experimental units are allocated to l^{th} treatment, for $l = 1, 2$, and $n_1 + n_2 = n$. The design matrix $\mathcal{U}$ may be written as

$$\mathcal{U} = \begin{pmatrix} J_{n_1} & 0 & x_1 \\ 0 & J_{n_2} & x_2 \end{pmatrix}, \tag{2.5}$$

where J_{n_l} is a n_l-dimensional vector with all elements as unity, corresponding to the parameter μ_l, and x_l are $n_l \times p$ matrices of the experimental units with $p-$covariates, which are allocated to the l^{th} treatment, $l = 1, 2$. For obtaining an optimal or near-optimal allocation design, we consider the information matrix $(\mathcal{U}^T W(\theta)\mathcal{U})$ or dispersion matrix $(\mathcal{U}^T W(\theta)\mathcal{U})^{-1}$, which depends upon the parameters. Let us assume that the parameters in θ, given by the μ_l's and the β_j's, independently follow the probability density functions $\pi(\mu_l)$ and $\pi(\beta_j)$, respectively, for $l = 1, 2$, $j = 1, 2, \cdots, p$. In this set-up, EW$D-$ and EW$A-$ optimality are obtained by minimizing, respectively, the determinant and trace of the estimated dispersion matrices with respect to priors $\pi(\mu_l)$ and $\pi(\beta_j)$, which are given by

$$det\ \{\int_{\mu_1}\int_{\mu_2}\int_{\beta_1}\cdots\int_{\beta_p} \mathcal{U}^T W(\theta)\mathcal{U}\ \pi(\mu_l)\pi(\beta_j)\ d\mu_l\ d\beta_j\}^{-1}, \tag{2.6}$$

and

$$trace\ \{\int_{\mu_1}\int_{\mu_2}\int_{\beta_1}\cdots\int_{\beta_p} \mathcal{U}^T W(\theta)\mathcal{U}\ \pi(\mu_l)\pi(\beta_j)\ d\mu_l\ d\beta_j\}^{-1}. \tag{2.7}$$

The information matrix $\mathcal{U}^T W(\theta)\mathcal{U}$ may be partitioned as

$$\mathcal{U}^T W(\theta)\mathcal{U} = \mathcal{I}(\theta) \quad = \begin{pmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{21} & \mathcal{I}_{22} \end{pmatrix}, \tag{2.8}$$

where $\mathcal{I}_{21} = \mathcal{I}_{12}^T$ and the individual elements may be obtained through the similar EW approach.

To obtain an optimal allocation design for the experimental units with known covariates into several treatment groups an efficient iterative search algorithm, called Hybrid Variable Neighborhood Search (VNS) algorithm, has been proposed in Hore et al. (2014, 2016). Hybrid VNS algorithm is a modification of the VNS algorithm, originally proposed by Hansen and Mladenovi̋c (1999), and has been developed in Hore et al. (2014, 2016, 2018), by introducing a single or multi-layered neighborhood structure with an in-built stochastic approach to find an optimal or near-optimal solution. In this study, the Hybrid VNS algorithm has been applied to obtain an optimal or near-optimal allocation design with regard to various $D-$ and $A-$optimality criteria under the GLM set-up.

3. Optimality under Poisson Regression Model

Poisson regression model is suitable when the response is a count variable and the mean response is the same as the corresponding variance. From equations (2.1)-(2.3), the mean of the response variable, i.e. $E(Y)$, for this model may be expressed as function of covariates $\mathcal{U}$ and the parameters θ with the obvious restriction of mean being positive. Thus, the response variable corresponds to the i^{th} experimental unit may be written as

$$E(Y_i) = exp(u_i^T \theta), \text{ for } i = 1, \cdots, n, \tag{3.1}$$

such that the logarithm of the mean function under such model is a linear function. The matrices in (2.5) may be written as $(n_l \times p)$ matrices $x_l = ((x_{li}^T))$, where $x_{li}^T = (x_{li1}, x_{li2}, ..., x_{lip})$ is a p-vector of covariates associated with the i^{th} unit, for $i = 1, \cdots, n_l$, and $l = 1, 2$. Thus, a closed form expression of the information matrix $\mathcal{I}(\theta)$ is available, in which the sub-matrices $\mathcal{I}_{11}, \mathcal{I}_{12}$ and $\mathcal{I}_{22}$ of $\mathcal{I}(\theta)$, as in (2.8), are given by $\mathcal{I}_{11} = diag\ (\sum_{i=1}^{n_1} e^{(\mu_1 + x_{1i}^T\beta)},\ \sum_{i=1}^{n_2} e^{(\mu_2 + x_{2i}^T\beta)})$, a 2×2 diagonal matrix,

$$\mathcal{I}_{12} \quad = \quad \begin{pmatrix} \sum_{i=1}^{n_1} x_{1i1} e^{(\mu_1 + x_{1i}^T\beta)} & ... & \sum_{i=1}^{n_1} x_{1ip} e^{(\mu_1 + x_{1i}^T\beta)} \\ \sum_{i=1}^{n_2} x_{2i1} e^{(\mu_2 + x_{2i}^T\beta)} & ... & \sum_{i=1}^{n_2} x_{2ip} e^{(\mu_2 + x_{2i}^T\beta)} \end{pmatrix},$$

a $2 \times p$ matrix, and

$$\mathcal{I}_{22} \quad = \quad \begin{pmatrix} \sum_{l=1}^{2}(\sum_{i=1}^{n_l} x_{li1}^2 e^{(\mu_l + x_{li}^T\beta)}) & ... & \sum_{l=1}^{2}(\sum_{i=1}^{n_l} x_{li1} x_{lip} e^{(\mu_l + x_{li}^T\beta)}) \\ \vdots & \vdots & \vdots \\ \sum_{l=1}^{2}(\sum_{i=1}^{n_l} x_{li1} x_{lip} e^{(\mu_l + x_{li}^T\beta)}) & ... & \sum_{l=1}^{2}(\sum_{i=1}^{n_l} x_{lip}^2 e^{(\mu_l + x_{li}^T\beta)}) \end{pmatrix},$$

a $p \times p$ matrix. We can easily evaluate the expectation of each element of the information matrix $\mathcal{I}(\theta)$ with respect to $\pi(\mu_l)$ and $\pi(\beta_j)$ as given, in general, by

$$\begin{aligned} & \int_{\mu_l} \int_{\beta_1} \cdots \int_{\beta_p} v_{lij}\ e^{(\mu_l + x_{li}^T\beta)} \pi(\mu_l)\ \pi(\beta_1) \cdots \pi(\beta_p)\ d\mu_l\ d\beta_1 \cdots d\beta_p \\ & = v_{lij}\ \{\int_{\mu_l} e^{\mu_l} \pi(\mu_l)\ d\mu_l\} \prod_{j=1}^{p} \{\int_{\beta_j} e^{\beta_j x_{lij}} \pi(\beta_j)\ d\beta_j\}, \\ & = v_{lij}\ M_{\mu_l}(1) \prod_{j=1}^{p} \{M_{\beta_j}(x_{lij})\}, \end{aligned} \tag{3.2}$$

where $M_{\mu_l}(1)$ is the moment generating function (m.g.f) of μ_l at 1, and $M_{\beta_j}(x_{lij})$ is the m.g.f of β_j at known covariate value x_{lij}, for $l = 1, 2,\ j = 1, 2, \cdots, p,\ i = 1, \cdots, n_l$, and v_{lij}

takes values 1, or x_{lij} or x_{lij}^2, or $x_{lij}x_{lij'}$, for fixed $j' \neq j$, as the case may be, for $l = 1, 2$, $j \neq j' = 1, 2, \cdots, p$ and $i = 1, \cdots, n_l$. Availability of such closed form expression (3.2) for the EW-version of the dispersion matrix, $E_\theta Var(\hat{\theta})$, alleviates the need for simulation method to obtain the same. In this work, two different sets of prior models for the parameters are assumed and the corresponding m.g.f's of these parameters are given in Table 2. However, other independent priors for the parameters may also be assumed.

Table 2. Moment generating function for assumed prior distributions.

Distribution	$\pi(\mu_l)$	$\pi(\beta_j)$	$M_{\mu_l}(1)$	$M_{\beta_j}(x_{lij})$
Normal	$N(\xi_{\mu_l}, \gamma_{\mu_l}^2)$	$N(\xi_{\beta_j}, \gamma_{\beta_j}^2)$	$e^{\xi_{\mu_l} + \frac{\gamma_{\mu_l}^2}{2}}$	$e^{x_{lij}\xi_{\beta_j} + \frac{x_{lij}^2 \gamma_{\beta_j}^2}{2}}$
Uniform	$U(a_{\mu_l}, b_{\mu_l})$	$U(a_{\beta_j}, b_{\beta_j})$	$\frac{e^{(b_{\mu_l} - a_{\mu_l})}}{(b_{\mu_l} - a_{\mu_l})}$	$\frac{e^{x_{lij}(b_{\beta_j} - a_{\beta_j})}}{(b_{\beta_j} - a_{\beta_j})}$

For small number of experimental units, say for $n \leq 10$, one can, as before (Hore et al.; 2014, 2016), compare the optimal allocation design obtained by the proposed Hybrid VNS algorithm with the one obtained through the actual search. However, for large or moderately large number of experimental units with single or multiple covariates, finding the exact optimal allocation is computationally intractable and one has to resort to the proposed Hybrid VNS algorithm. As before, the efficiency comparison is carried out with respect to the random allocation design. However, closed form expression for $E_\theta Var(\hat{\theta})$ is not available for the other regression models given in Table 1. To circumvent this problem, one may consider a suitable simulation method to approximate $E_\theta Var(\hat{\theta})$. Nevertheless, the proposed Hybrid VNS algorithm can be used for optimization over all possible allocation designs.

The performance of the proposed Hybrid VNS algorithm with respect to the exact optimal search has been carried out for small number of experimental units ($n = 10$) of a single covariate with regard to $D-$ and $A-$optimality. The covariate has been generated from different probability distributions, like Uniform distribution with parameters 0 and 1, denoted by $U[0, 1]$, Normal distribution with mean 0 and variance 5, i.e. $N(0, 5)$ and Exponential distribution with mean 25 (or rate 0.4), $E(0.4)$. For known covariate values and a given allocation design, we can easily calculate the information matrix and the values of the objective functions corresponding to $D^{(1)}-$ and $A^{(1)}-$optimality through the moment generating functions of the prior distribution of the parameters, as given in expression (3.2). Hence, the exact optimal allocation design may be obtained among the total search of $(2^9 - 1)$ nontrivial allocation combinations (Hore et al., 2014). The efficiency is defined as the ratio of the values of the objective functions, denoted by $V(\cdot)$, say, for the exact optimal design (α^e), and the allocation design obtained by the proposed method (α^P), i.e. $\frac{V(\alpha^e)}{V(\alpha^P)}$. The average of this efficiency value over 1000 simulated sets of n covariate values are reported in Table 3. It is observed that the mean efficiency never falls below 90% whatever be the choice of prior and covariate distributions.

With large or moderately large number of experimental units ($n = 50$ or 100), finding the exact optimal allocation design searching through $(2^{49}-1)$ or $(2^{99}-1)$ allocation combinations, respectively, is computationally intractable. For still larger experimental units, the computational burden is enormous and nearly impossible. In such situations, by following what has been done in Hore et al. (2014) under LM set-up, the efficiency of the near-optimal design obtained by the proposed Hybrid VNS algorithm (α^P) may be compared with the randomized allocation design (α^R) with regard to $D-$ and $A-$optimality through simulation experiments where the corresponding efficiency is defined as $\frac{V(\alpha^P)}{V(\alpha^R)}$. The covariates are generated from a bivariate normal (BVN) distribution for $D^{(\nu)}$ and $A^{(\nu)}-$optimality, $\nu = 1, 2, 3$. The simulation study has been carried out 1000 times in each case and the results are summarized in Tables 4 and 5 for $D^{(\nu)}$ and $A^{(\nu)}-$optimality, respectively. Although we have considered the number of experimental units to be 50 and 100 for illustration purpose, the algorithm may be implemented to accommodate still larger number of units requiring only additional but manageable computer time, that can be executed in any standard laptop or desktop with modest computational facility. As we have assumed that the covariate values corresponding to the experimental units are given and prefixed before the experiment is conducted, throughout the paper we have considered 50 or 100 number of experimental units with corresponding covariate values being generated from $BVN(10, 5, 4, 5, 0.5)$. Here $\theta = (\mu_1, \mu_2, \beta_1, \beta_2)$ is the vector of parameters of interest having corresponding prior distribution $\pi(\theta)$. Three different kinds of prior distribution of the parameters are considered to reflect different patterns in them, viz. symmetric, flat or heavy tailed. Consequently, we have chosen the priors from (1) Normal distribution with parameters μ and σ^2, i.e. $N(\mu, \sigma^2)$, (2) Uniform distribution with parameters a, b, i.e. $U[a, b]$ and (3) Laplace distribution with location parameter μ and scale parameter b, denoted by $Lp(\mu, b)$ with density

$$f(x\ ;\ \mu,\ b) = \frac{1}{2b}\, e^{-\frac{|x-\mu|}{b}}\ ; \quad -\infty < x < \infty\ , \quad -\infty < \mu < \infty\ , \quad b > 0\ .$$

These prior distributions of the parameters along with the values of the hyper-parameters are chosen arbitrarily, only for the illustration purpose. It has been observed that the VNS algorithm performs uniformly better than the random allocation design and sometimes with more than two fold increase in the efficiency value. It has been also noted that the efficiency is never lesser than unity even for prior with large variance.

Table 3. Mean efficiency of the $D^{(1)}-$ and $A^{(1)}-$optimal allocation designs obtained by the proposed algorithm with respect to the exact allocation over 1000 simulations (range in square brackets) for known single covariate of $n = 10$ units from different covariate distributions under Poisson regression model with prior distributions $\pi(\theta)$.

$\pi(\theta)$	Covariate Distribution	$D^{(1)}$	$A^{(1)}$
$\mu_1 \sim N(0.5, 1)$,	U[0,1]	0.9937 [0.9714, 1]	0.9915 [0.9683, 1]
$\mu_2 \sim N(0.75, 2)$,	N(0,5)	0.9957 [0.9446, 1]	0.9938 [0.9329, 1]
and $\beta \sim N(1, 1.5)$	E(0.4)	0.9962 [0.9510, 1]	0.9941 [0.9418 ,1]
$\mu_1 \sim N(0.5, 100)$,	U[0,1]	0.9418 [0.8912, 1]	0.9439 [0.8861, 1]
$\mu_2 \sim N(0.75, 2)$,	N(0,5)	0.9515 [0.9061, 1]	0.9511 [0.8983, 1]
and $\beta \sim N(1, 1.5)$	E(0.4)	0.9409 [0.8943, 1]	0.9413 [0.8937 ,1]
$\mu_1 \sim N(0, 10000)$,	U[0,1]	0.9083 [0.8651, 1]	0.9079 [0.8619, 1]
$\mu_2 \sim N(10, 500)$,	N(0,5)	0.9042 [0.8735, 1]	0.9037 [0.8724, 1]
and $\beta \sim U[-250, 250]$	E(0.4)	0.9038 [0.8653, 1]	0.9023 [0.8618 ,1]

Table 4. Mean efficiency of the $D-$optimal allocation designs obtained by the proposed algorithm with respect to the random allocation over 1000 simulations (range in square brackets) for known covariates generated from BVN (10,5,4,5,0.5) of n experimental units under Poisson regression model and prior distribution $\pi(\theta)$.

$\pi(\theta)$	n	$D^{(1)}$	$D^{(2)}$	$D^{(3)}$
$\mu_1 \sim N(1, 4), \mu_2 \sim N(2, 5)$	50	1.493 [1.092, 2.213]	1.459 [1.081, 2.103]	1.474 [1.083, 1.768]
$\beta_1 \sim U[-2, 2], \beta_2 \sim U[0, 5]$	100	1.431 [1.071, 1.834]	1.382 [1.066, 1.513]	1.386 [1.063, 1.627]
$\mu_1 \sim N(1, 4000), \mu_2 \sim N(2, 500)$	50	1.268 [1.083, 1.673]	1.242 [1.073, 1.642]	1.249 [1.079, 1.638]
$\beta_1 \sim Lp(0, 100), \beta_2 \sim U[-250, 250]$	100	1.249 [1.067, 1.604]	1.231 [1.059, 1.513]	1.241 [1.059, 1.587]

Table 5. Mean efficiency of the $A-$optimal allocation designs obtained by the proposed algorithm with respect to the random allocation over 1000 simulations (range in square brackets) for known covariates generated from BVN (10,5,4,5,0.5) of n experimental units under Poisson regression model and prior distribution $\pi(\theta)$.

$\pi(\theta)$	n	$A^{(1)} = A^{(2)}$	$A^{(3)}$
$\mu_1 \sim N(1,4), \mu_2 \sim N(2,5)$	50	1.427 [1.077, 1.801]	1.419 [1.071, 1.784]
$\beta_1 \sim U[-2,2], \beta_2 \sim U[0,5]$	100	1.378 [1.083, 1.764]	1.373 [1.077, 1.759]
$\mu_1 \sim N(1,4000), \mu_2 \sim N(2,500)$	50	1.253 [1.053, 1.662]	1.233 [1.048, 1.614]
$\beta_1 \sim Lp(0,100), \beta_2 \sim U[-250,250]$	100	1.219 [1.048, 1.537]	1.209 [1.041, 1.518]

4. Logistic regression model

In the binary logistic regression set up, we assume that the response variable is binary. Hence, from equations (2.1)-(2.3), the response variable corresponding to the i^{th} experimental unit, Y_i, may be modeled as

$$E(Y_i) = \frac{exp(u_i^T\theta)}{(1 + exp(u_i^T\theta))}, \text{ for } i = 1, \cdots, n, \tag{4.1}$$

which is positive but always less than the unity. The above model may be rewritten as follows

$$log\{\frac{E(Y_i)}{1 - E(Y_i)}\} = u_i^T\theta, \text{ for } i = 1, \cdots, n;$$

which is called the logit function and the equation (4.1) is called the logistic regression model.

Unlike the Poisson regression model, the elements of the information matrix under logistic regression model do not have any closed form expression. Consequently, the simulation approach provides an alternative for obtaining the integrals under expectations. The performances of the proposed algorithm over the randomized method with regard to various $D-$ and $A-$optimality criteria have been explored and are reported in Tables 6 and 7, respectively. We take the same covariate values and same prior distributions for the parameters as in Section 3. Here also, the proposed algorithm performs uniformly better than the randomized allocation design and sometimes found to be 2 times more efficient than the random allocation with the efficiency value never falling below unity even for prior distributions with larger variance. The simulation studies discussed here and in previous section have also been carried out for other covariate values in bivariate set-up and other prior distributions with different choices of the hyper-parameters. The findings are in general similar.

Table 6. Mean efficiency of the $D-$optimal allocation designs obtained by the proposed algorithm with respect to the random allocation over 1000 simulations (range in square brackets) for known covariates generated from BVN (10,5,4,5,0.5) of n experimental units under logistic regression model and prior distribution $\pi(\theta)$.

$\pi(\theta)$	n	$D^{(1)}$	$D^{(2)}$	$D^{(3)}$
$\mu_1 \sim N(1,4), \mu_2 \sim N(2,5)$	50	1.387 [1.084, 1.829]	1.372 [1.076, 1.761]	1.374 [1.073, 1.752]
$\beta_1 \sim U[-2,2], \beta_2 \sim U[0,5]$	100	1.347 [1.073, 1.728]	1.339 [1.067, 1.713]	1.335 [1.063, 1.7 27]
$\mu_1 \sim N(1,4000), \mu_2 \sim N(2,500)$	50	1.252 [1.069, 1.623]	1.237 [1.065, 1.597]	1.239 [1.069, 1.588]
$\beta_1 \sim Lp(0,100), \beta_2 \sim U[-250,250]$	100	1.227 [1.056, 1.574]	1.221 [1.052, 1.539]	1.222 [1.058, 1.537]

Table 7. Mean efficiency of the $A-$optimal allocation designs obtained by the proposed algorithm with respect to the random allocation over 1000 simulations (range in square brackets) for known covariates generated from BVN (10,5,4,5,0.5) of n experimental units under logistic regression model and prior distribution $\pi(\theta)$.

$\pi(\theta)$	n	$A^{(1)} = A^{(2)}$	$A^{(3)}$
$\mu_1 \sim N(1,4), \mu_2 \sim N(2,5)$	50	1.316 [1.069, 1.718]	1.308 [1.061, 1.684]
$\beta_1 \sim U[-2,2], \beta_2 \sim U[0,5]$	100	1.278 [1.063, 1.713]	1.264 [1.065, 1.659]
$\mu_1 \sim N(1,4000), \mu_2 \sim N(2,500)$	50	1.227 [1.058, 1.587]	1.218 [1.054, 1.543]
$\beta_1 \sim Lp(0,100), \beta_2 \sim U[-250,250]$	100	1.223 [1.043, 1.388]	1.216 [1.038, 1.367]

5. Imbalance in Allocation Design

In classical linear ANCOVA model, balancing is defined through equality of covariate means corresponding to various treatment groups and such balancing ensures widely used $D-$ and $A-$optimality (Shah and Sinha, 1989). It has also been analytically established by Hore et al. (2017) that for the allocation of experimental units with known categorical covariates to two treatments, 'balancing' is also the necessary condition towards achieving $D-$ and $A-$optimality. Interestingly, for continuous covariates, it has been empirically observed (Hore et al., 2014, 2016) that optimality of various kinds result in near balance situation. Therefore, optimality and balancing may be seen as being approximately synonymous in allocation design under LM set-up. In real life, for a high dimensional problem with large number of experimental units, attaining exact balance is impractical and nearly impossible and hence one has to be content with an approximate balancing. Consequently, the proposed optimal allocation procedure has got immense relevance. Our next objective is to study the status of balancing with regard to various kinds of optimality if we move from LM to GLM set-up. Consider the squared Mahalanobis distance (Mahalanobis, 1936) or Mahalanobis imbalance metric (Imbens and Rubin, 2015, p 342), say M, as an imbalance measure with p known continuous covariates, given by

$$M \;=\; (\overline{x_{(1)}} - \overline{x_{(2)}})^T \left[\frac{n_1\hat{\Sigma}_1 + n_2\hat{\Sigma}_2}{n_1+n_2}\right]^{-1} (\overline{x_{(1)}} - \overline{x_{(2)}}), \tag{5.1}$$

where $\overline{x_{(l)}} = \frac{1}{n_l}\sum\limits_{i=1}^{n_l} x_{li}$, is the $p-$component column vector of covariate means over the units allocated to the l^{th} treatment group with sample sizes n_l and $\hat{\Sigma}_l = \frac{1}{n_l}\sum\limits_{i=1}^{n_l}(x_{li}-\bar{x}_l)(x_{li}-\bar{x}_l)^T$ represents the sample covariance matrix of the known covariates in the lth treatment group, for $l = 1, 2$.

The nature of this imbalance measure for the optimal allocation design with regard to several optimality criteria for known continuous covariates while comparing two treatments under both the LM and the and GLM set-up are considered through an extensive simulation study. In all such exercises, we have considered the problem of allocating 100 experimental units, whose covariates, as before, are generated from $BVN = (10, 5, 4, 5, 0.5)$, and their values are kept fixed. The prior distribution is $\pi(\mu_1, \mu_2, \beta_1, \beta_2)$, where $\mu_1 \sim N(1, 4)$, $\mu_2 \sim N(2, 5)$, $\beta_1 \sim U[-2, 2]$ and $\beta_2 \sim U[0, 5]$ and all are independent, as in upper panel of Tables 4 to 7. For a progressive assessment of the nature of balance over the successive iterations of the Hybrid VNS algorithm with these 100 experimental units with fixed covariate values, we have computed the Mahalanobis imbalance metric M of (5.1) corresponding to the allocation design, obtained at each iteration of the VNS algorithm. The paths of such values over different iterations for various optimal allocation schemes (See Figures 1 and 2) seem to stabilize after few iterations and converge to a value, which might be taken as the value of the Mahalanobis imbalance metric for that particular optimal allocation scheme. In Figure 1, we have considered different kinds of $D-$ and $A-$optimality for GLM (Poisson) set-up along with the usual D- and A-optimality for the LM. In Figure 2, the same optimality criteria for GLM (logistic) and the LM set-up are reported. It is apparent from the figures that, under the LM formulation, $D-$ and $A-$optimal allocation designs result in the zero value for the Mahalanobis imbalance metric. But, under the GLM set-up, the values of this metric M corresponding to the optimal designs in various senses are far from zero, although there is evidence of existence of sub-optimal designs in earlier iterations with smaller values of this imbalance metric M. This suggests that a more balanced design may be obtained by sacrificing optimality to some extent. In other words, it may be reasonable to constrain the values of M and then look for optimality.

Figure 1: Comparison of convergence of the Mahalanobis Distance over various iterations corresponding to $D-$ and $A-$optimality under Poisson GLM and linear ANCOVA model.

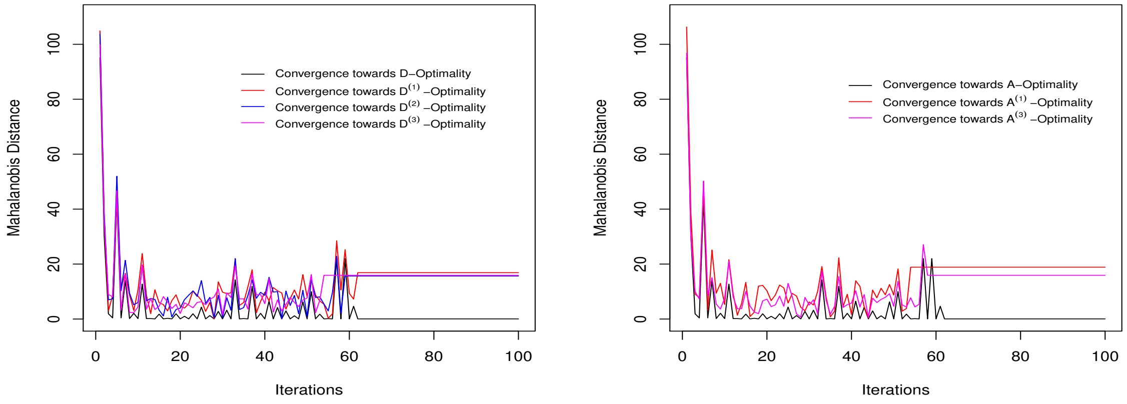
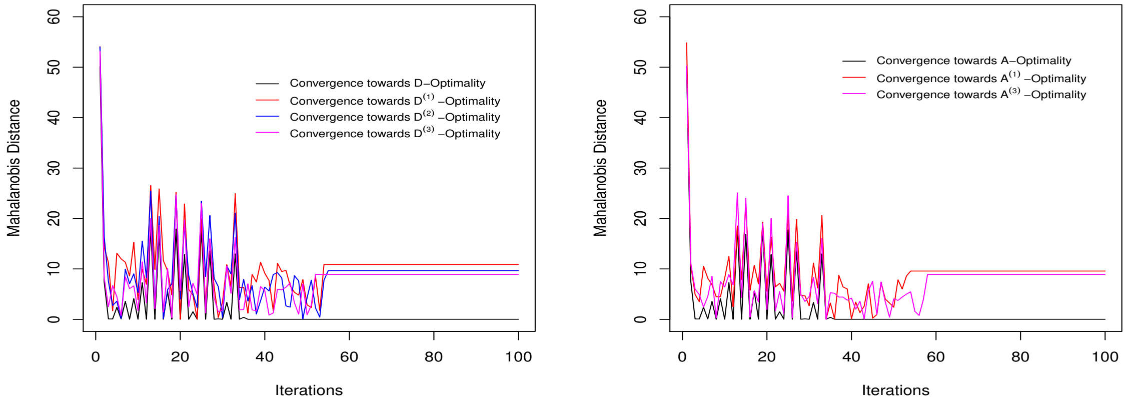


Figure 2: Comparison of convergence of the Mahalanobis Distance over various iterations corresponding to $D-$ and $A-$optimality under logistic GLM and linear ANCOVA model.

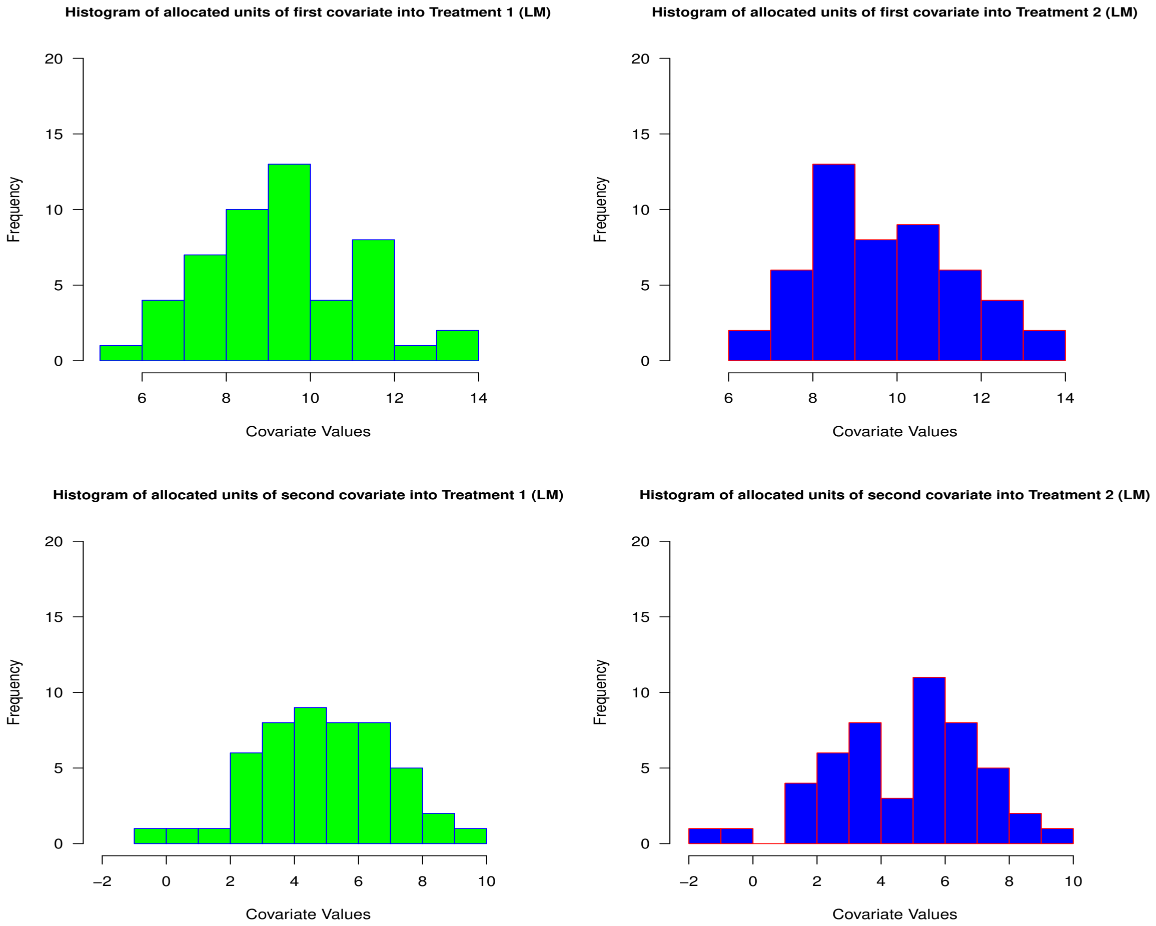


Graphical display of distributional studies of the two covariates in the two treatment groups through histograms corresponding to the $D-$optimal allocation design under both the LM and the GLM formulations with Poisson and logistic regression models (with same priors as before) are shown in Figures 3, 4 and 5, respectively. For the three different modeling scenarios, the values of M corresponding to the different D-optimal allocation designs and the associated p-values for testing equality of covariate means in the two treatment groups using the $\chi^2_{(2)}$ statistic (Morgan and Rubin, 2012) are calculated. Generally, M is the Hotelling T^2 statistic, but here M follows a $\chi^2_{(p)}$ distribution because $p-$covariate values are considered as fixed and known (Mardia, Kent and Bibby, 1980, p 62; Morgan and Rubin, 2012). For the LM set-up in Figure 3, the value of the metric is 0.00136 and the corresponding p-value is 0.99932 suggesting minimum imbalance and near equality of covariate means in the two groups for the $D^{(1)}-$optimal allocation design. For the two GLM cases in Figures 4 and 5, the values of M corresponding to the $D^{(1)}$-optimal designs are much higher and the $p-$values are smaller than 0.05 with evidences against equality of covariate means in the two treatment groups. Thus, under optimal allocation in various senses, the covariate distributions are more or less balanced over the two treatment groups in the LM set-up, while they are far from balanced in the GLM cases, for the same set of known covariate values associated with the experimental units. We have carried out this exercise with several other sets of bivariate covariates simulated from other bivariate distributions with extreme choices of hyper parameters, as in the lower panel of Tables 4 to 7, and the findings are more or less similar.

6. A compromise between Balance and Optimality

In view of the empirical results of the previous section regarding the close relationship between covariate balance and optimality under LM set-up, and the absence of it under various GLM set-up, it may be of interest to obtain an allocation scheme under the GLM set-up that maintains a compromise between optimality and imbalance. With this objective, as indicated in the previous section, a constrained optimality problem has been suggested where optimality in any sense might be attempted among the class of allocation designs with an upper limit on the Mahalanobis imbalance metric M computed using (5.1) on the basis of the covariate values in the two treatment groups. We consider a pre-fixed upper limit, say a, for the imbalance metric M and a near-optimal allocation design among the class of all allocation designs with $M \leq a$, is obtained through a suitably constrained VNS

Figure 3: Histograms of covariates for the 100 experimental units allocated into two treatment groups with regard to $D-$optimality under the LM set-up (value of the Mahalanobis imbalance metric is 0.00136 with the p-value = 0.99932).



Histogram of allocated units of first covariate into Treatment 1 (LM)

Histogram of allocated units of first covariate into Treatment 2 (LM)

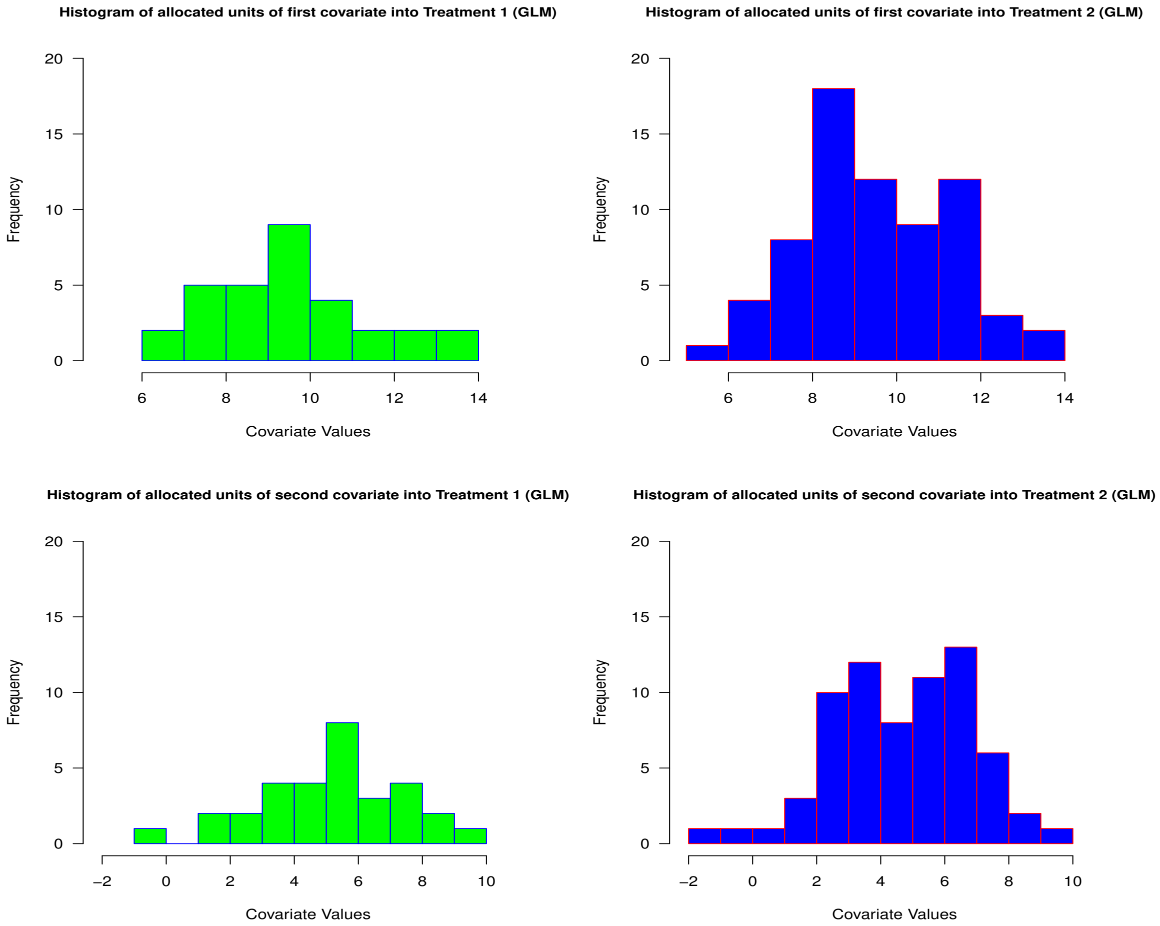
Histogram of allocated units of second covariate into Treatment 1 (LM)

Histogram of allocated units of second covariate into Treatment 2 (LM)

algorithm. As in the re-randomization approach of Morgan and Rubin (2012), the method exploits the proposed Hybrid VNS algorithm subject to the rejection of any allocation design whenever $M > a$. Noting that the Mahalanobis imbalance metric M follows approximately a χ^2 distribution with p degrees of freedom for p continuous covariates (Morgan and Rubin, 2012), we take the lower 100δth percentile of $\chi^2_{(p)}$ distribution as a, the upper limit of the imbalance metric M, representing a flexible and acceptable benchmark balance for some small $\delta > 0$. In other words, we search for optimality among that class of allocation designs which have the value of the Mahalanobis imbalance metric less than the 100δth percentile point of the $\chi^2_{(p)}$ distribution. For the demonstration purpose, we take δ to be 0.1.

The covariate distributions corresponding to the unconstrained optimal allocation for Poisson and logistic models are reported in Figures 4 and 5, respectively, on the basis of the

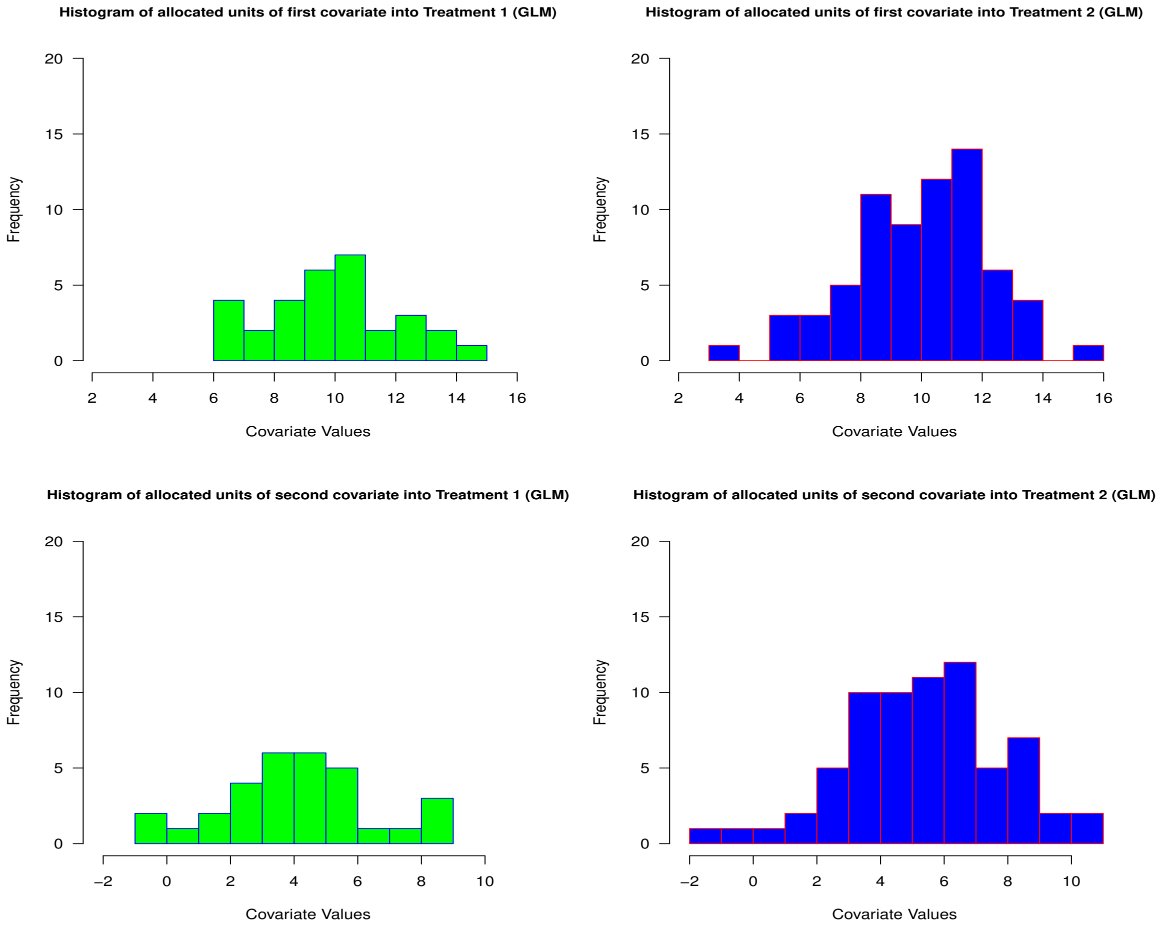
Figure 4: Histograms of covariates for the 100 experimental units allocated into two treatment groups with regard to $D^{(1)}-$optimality under the Poisson GLM set-up (value of the Mahalanobis imbalance metric is 6.42645 with the p-value = 0.04023).



same set of covariate values, as has been done for the LM in Figure 3. Similar covariate distributions, with the same prior distributions as considered in Section 5, corresponding to the constrained problem are reported in Figures 6 and 7 for Poisson and logistic models, respectively, where the upper limit of the Mahalanobis imbalance metric is kept as the 10th percentile of the $\chi^2_{(2)}$ distribution (i.e. 0.21072). It is apparent from Figures 6 and 7 that balance in the covariate distribution is achieved to a larger extent in comparison with those reported in Figures 4 and 5 with some sacrifice in optimality as shown through the efficiency study in Tables 8 and 9.

The efficiency of the constrained optimal allocation design α^M, say, has been investigated through simulation studies over 1000 repetitions, with respect to the $D^{(\nu)}$ and $A^{(\nu)}-$optimal allocation designs $\alpha^{D^{(\nu)}}$ and $\alpha^{A^{(\nu)}}$, respectively, for $\nu = 1, 2, 3$. For Poisson and logistic

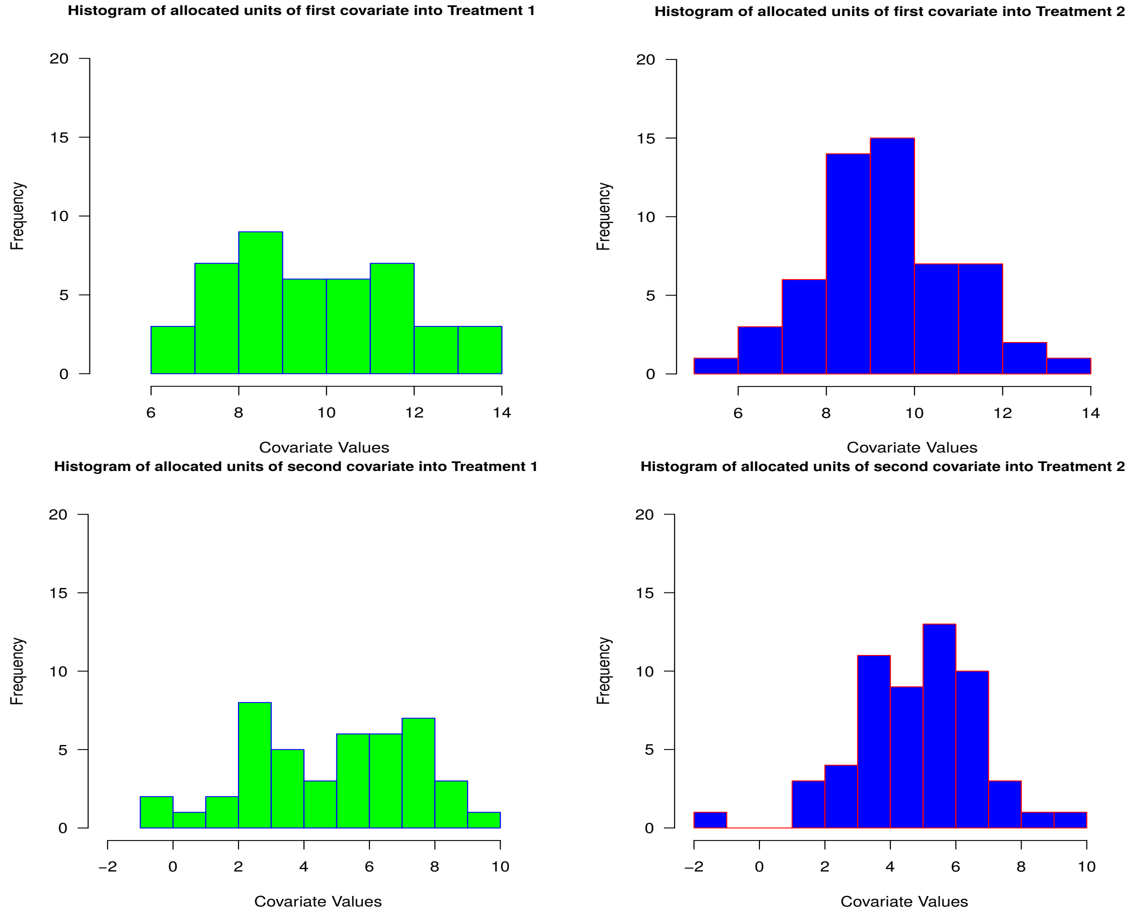
Figure 5: Histograms of covariates for the 100 experimental units allocated into two treatment groups with regard to $D^{(1)}-$optimality under the binary logistic GLM set-up (value of the Mahalanobis imbalance metric is 7.58326 with the p-value = 0.02252).



models, the results are reported in Tables 8 and 9, respectively, while keeping the covariate values and the prior distributions same as before. It is to be observed that measures of efficiency and imbalance metric have an inverse relationship. As for example, the mean efficiency is found to be as low as 40% if the extent of imbalance is restricted to the 5th percentile, while it can be as high as 77% if the extent of imbalance is relaxed up to the 20th percentile cutoff.

Table 8. Mean efficiency of the constrained optimal design with the threshold at 100δth percentile with respect to $D-$ and $A-$optimal allocation designs over 1000 simulations for allocating 100 experimental units under Poisson GLM and the same prior distribution.

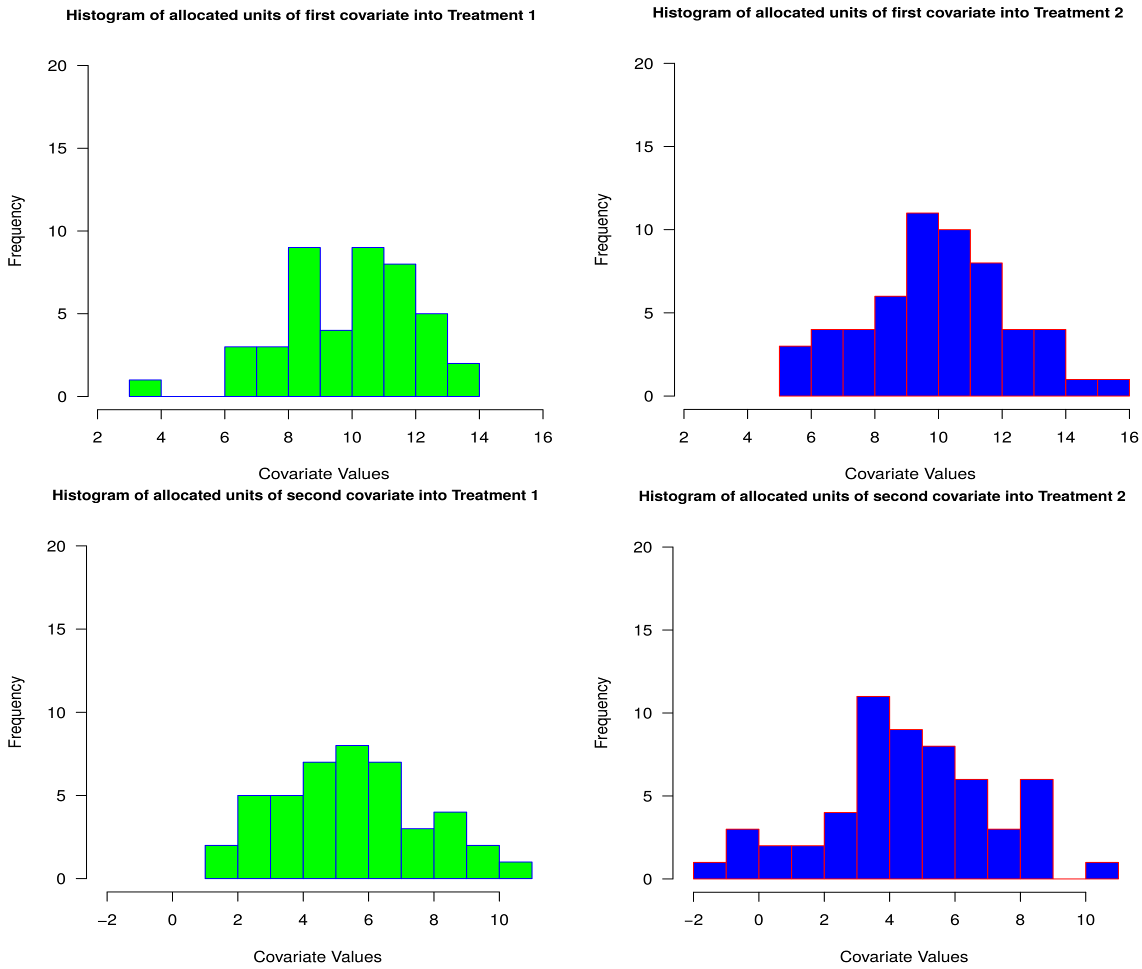
Figure 6: Histogram of the covariates corresponding to the 100 experimental units to be allocated into two groups with regard to $D^{(1)}-$optimality under the Poisson GLM set-up subject to a threshold value of 0.21072 for balance (value of the Mahalanobis imbalance metric is 0.20928 with the p-value = 0.90065).



δ	$D^{(1)}$	$D^{(2)}$	$D^{(3)}$	$A^{(1)}$	$A^{(3)}$
0.20	0.7732	0.7586	0.7608	0.7324	0.7196
0.10	0.5739	0.5542	0.5618	0.5351	0.5237
0.05	0.4842	0.4527	0.4719	0.4122	0.4103

Table 9. Mean efficiency of the constrained optimal design with the threshold at 100δth percentile with respect to $D-$ and $A-$optimal allocation designs over 1000 simulations for allocating 100 experimental units under logistic GLM and the same prior distribution.

Figure 7: Histogram of the covariates corresponding to the 100 experimental units to be allocated into two groups with regard to $D^{(1)}-$optimality under the logistic GLM set-up subject to a threshold value of 0.21072 for balance (value of the Mahalanobis imbalance metric is 0.20516 with the p-value $= 0.90250$).



δ	$D^{(1)}$	$D^{(2)}$	$D^{(3)}$	$A^{(1)}$	$A^{(3)}$
0.20	0.7283	0.7069	0.7212	0.6816	0.6762
0.10	0.5506	0.5233	0.5421	0.5108	0.5037
0.05	0.4319	0.4106	0.4192	0.3928	0.3912

7. Analysis of Real Life Data

The efficacy of the proposed Hybrid VNS algorithm has been illustrated with the help of a real data set, already considered by Hore et al. (2016) in the LM set-up. A two-years-long study, carried out on agents to reduce dental caries in 69 female children, has been reported in Quade (1982). Three treatments, namely stannous fluoride (SF), acid-phosphate fluoride (APF), and the distilled water (W) as placebo, are used to study the effectiveness of SF and APF as caries-reducing agents in children. The number of decayed, missing, or filled teeth (DMFT) is reported before (B) and after (A) the study where the response (Y) is the difference between A and B, i.e., (A-B). The initial DMFT count (B) and the age (X in years) at the beginning of the study are considered as covariates. Note that the response Y is a count variable and the Poisson GLM is the appropriate set-up for that. To make it a two-treatment problem, for the sake of illustration, we ignore the experimental units allocated to W and consider only those units with non-negative responses which are allocated to either SF or APF. The reduced data set with 48 female children are considered which are allocated to the two treatment groups SF and APF. Assuming the same independent prior distributions for the parameter $\theta = (\mu_1, \mu_2, \beta_1, \beta_2) \sim N(1,4)$, $N(2,5)$, $U[-2,2]$, $U[0,5]$, we find that the efficiency of our algorithm with respect to random allocation design as 1.583(1.673), 1.746(1.752) and 1.767(1.766) for $D^{(1)} - (A^{(1)}-)$optimality, $D^{(2)} - (A^{(2)}-)$optimality and $D^{(3)-}(A^{(3)}-)$optimality, respectively, with Mahalanobis imbalance metric lying between 3.24125 to 5.83247. A constrained optimal allocation design is obtained through the proposed algorithm with the upper limit for imbalance kept at the 10th percentile of $\chi^2_{(2)}$ distribution (i.e. 0.21072). The efficiency values with respect to the corresponding optimal allocation design turns out to be 0.6342(0.6063), 0.6128(0.6237) and 0.6229(0.6019) for $D^{(1)} - (A^{(1)}-)$optimality, $D^{(2)} - (A^{(2)}-)$optimality and $D^{(3)-}(A^{(3)}-)$optimality, respectively.

It is to be noted that an optimal allocation design under the LM set-up, although easier to obtain, may be sub-optimal if it is more appropriate to assume the Poisson GLM set-up as the correct model. In order to study the extent of such sub-optimality, we have computed the efficiency values for such sub-optimal designs with regard to the optimal designs under the Poisson GLM set-up keeping the prior distributions same. These efficiency values are 0.6428(0.6231), 0.6049(0.6231) and 0.6156(0.6191) for $D^{(1)} - (A^{(1)}-)$optimality, $D^{(2)} - (A^{(2)}-)$optimality and $D^{(3)-}(A^{(3)}-)$optimality, respectively. Note the similarity between these efficiency values with those for the constrained designs reported earlier. This is not surprising since the optimal design under the LM set-up is found to result in minimum imbalance. On the other hand, assuming the LM to be the correct model, the efficiency values of the $D^{(1)} - (A^{(1)}-)$optimal designs under a Poisson GLM set-up, in comparison to the optimal designs under the LM formulation, are obtained as 0.6329(0.6217) for $D-(A-)$optimality. Thus, it appears that misspecification of the model (LM against GLM and vice versa) results in about 40% sacrifice of efficiency. Such comparison has also been done for other prior distributions with different sets of hyper-parameters and the general findings are similar.

8. Concluding Remarks

This paper deals with the allocation of a given set of experimental units with known covariate values into two treatment groups under the GLM set-up for the efficient estimation of 'treatment and covariate effects'. To achieve an optimal or near-optimal allocation design under different optimality criteria, the Hybrid VNS algorithm (Hore et al,. 2014, 2016) has been applied. Through simulation studies, it has been found that the proposed algorithm performs uniformly better than the randomized allocation. One can think of the corresponding D_s or A_s-optimality involving only the treatment effects requiring consideration of only the corresponding sub-matrix of $Var(\hat{\theta})$ in (2.4). The algorithm thereafter follows similarly. It has been observed that a near-optimal allocation design attains minimum Mahalanobis imbalance metric under the LM set-up, while it fails to attain such minimum imbalance under the GLM set-up. In view of this, for the GLM set-up, an upper limit for the imbalance metric has been proposed and the optimal allocation rule is obtained under such constraint. This ensures neither the global optimality nor the minimum imbalance, but strives a compromise between these two entities. As the balanced allocation with regard to observed covariates is a desirable criterion in causal inference (Rubin, 2008 ; Morgan and Rubin, 2012) and factorial design (Branson et al., 2016), such compromise might be useful and worth exploring. The work may be extended for multiple treatments and other modeling assumptions. Apart from such covariate balance one might be interested in proportional balance, or a scheme with uneven allocation ratio. In many clinical studies, for example, a preference is shown for the experimental treatment over the standard treatment with an allocation ratio 2:1 or even 3:1, in favor of the experimental treatment. In such context, a new measure of imbalance may be needed with the proportion of allocated experimental units to different treatments reflecting the required allocation ratio. A dual problem may also be suggested where the objective is to minimize the imbalance measure with a lower bound for efficiency with respect to some standard design e.g. randomized allocation rule.

The proposed algorithm may also be used to obtain the optimal robust allocation under few competing models, say logit and probit for binary response data, under some suitable priors for the parameter(s). Such model robustness approach is widely used in toxicological experiment under different non-linear models (Dette et al., 2012; Braess et al., 2013). These are being currently studied and will be reported afterwards.

References:

1. Abdelbasit, K.M. and R.L. Plackett (1983): Experimental design for binary data. *Journal of American Statistical Association*, **78**, 90-98.

2. Braess D, Dette H. (2013) : Optimal discriminating designs for several competing regression models. The Annals of Statistics, **41** (2): 897-922.

3. Branson, Z. Dasgupta, T. and Rubin, D.B. (2016) : Improving covariate balance in 2^k factorial designs via rerandomization with an application to a new york city department of education high school study. *The Annals of Applied Statistics*, **10** (4), 1958-1976.

4. Cartwright, H.V., Lindahl, R.L. and Bawden, J.W. (1968): Clinical findings on the effectiveness of stannous fluoride and acid phosphate fluoride as caries reducing agents in children. *Journal of Dentistry for Children*, **35**, 36-40.

5. Chernoff, H. (1953): Locally optimal designs for estimating parameters. *Ann. Math. Statist*, **24**, 586-602.

6. Dette H, Melas VB, Shpilev P. (2012) : T-optimal designs for discrimination between two polynomial models. The Annals of Statistics, 40 (1): 185-205.

7. Finney, D.J. (1978): *Statistical Methods in Biological Assay*, (third edition). Macmillan, New York.

8. Hansen, P. and Mladenovic ́, N. (1999) : An introduction to variable neighborhood search. in : S. Voss, S. Martello, I. Osman, and C. Roucairol, C. (Eds.), Metaheuristics: Advances and Trends in Local Search Paradigms for Optimization, Kluwer Academic Publishers, MA, USA, 433-458.

9. Hastings, W. K. (1970) : Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57** (1), 97-109.

10. Harville, D.A. (1974): Nearly optimal allocation of experimental units using observed covariate values. *Technometrics*, **16** (4), 589-599.

11. Hinkelmann, K. and Kempthorne, O. (2005): *Design and Analysis of Experiments*, Volume 1, New York: John Wiley and Sons Inc.

12. Hinkelmann, K. and Kempthorne, O. (2007): *Design and Analysis of Experiments*, Volume 2, New York: John Wiley and Sons Inc.

13. Hore, S., Dewanji, A. and Chatterjee, A. (2014) : Design issues related to allocation of experimental units with known covariates into two treatment groups. *Journal of Statistical Planning and Inference*, **155**, 117-126.

14. Hore, S., Dewanji, A. and Chatterjee, A. (2016) : On Optimal Allocation of Experimental Units with Known Covariates into Multiple Treatment Groups. *Calcutta Statistical Association Bulletin*, **68**(1 & 2), 69-81.

15. Hore, S., Dewanji, A., Chatterjee, A. (2017). Ensuring balance through optimal allocation of experimental units with known categorical covariates into two treatments. Technical Report No. ASU/2017/12, Dated : 26 July, 2017.

16. Hore, S., Chatterjee, A., Dewanji, A. (2018). Improving variable neighborhood search to solve the traveling salesman problem. *Applied Soft Computing*, **68** (July 2018), 83-91.

17. Hu,Y. and Hu,F. (2012a): Balancing treatment allocation over continuous covariates: a new imbalance measure for minimization. *Journal of Probability and Statistics*, DOI: 10.115/2012/842369.

18. Hu,Y. and Hu,F. (2012b): Asymptotic properties of covariate-adaptive randomization. *The Annals of Statistics* **40**(3), 1794 - 1815.

19. Imbens, G. W. and Rubin, D. B. (2015): Causal inference for statistics, social and biomedical sciences, An introduction, Cambridge University Press, New York.

20. Kalbfleisch, J.D. and Prentice, R.L. (2002) : *The Statistical Analysis of Failure Time Data*, Second Edition, New York: John Wiley and Sons Inc.

21. Kiefer, J. (1959): Optimum Experimental Designs, *Journal of Royal Statistical Society, Series B* **21** (2), 272-319.

22. MacKenzie, G. and Peng, D. (2014) : *Statistical Modelling in Biostatistics and Bioinformatics : Selected Papers.* Switzerland : Springer International Publishing.

23. Mahalanobis, P.C. (1936) : On the Generalized Distance in Statistics. *Proceedings of the National Institute of Sciences (Calcutta)* **2**, 49-55.

24. Mardia, K. V., Kent, J. T. and Bibby, J. M. (1980) : *Multivariate Analysis.* Academic Press, London.

25. McCullagh, P. and Nelder, J. A. (1989) : *Generalised Linear Models*, Second Edition, Boca Raton : Chapman and Hall / CRC.

26. Morgan, B. J. T. (1992) : Analysis of quantal response data. Springer : New York.

27. Morgan, K. L. and Rubin, D. B. (2012) : Rerandomization to improve covariates balance in experiments. *The Annals of Statistics*, **40** (2), 1263-1282.

28. Quade, D. (1982): Nonparametric analysis of covariance by matching. *Biometrics* **38** (3), 597-611.

29. Rosenberger W. F., Sverdlov O. (2008) : Handling covariates in the design of clinical trials. *Statistical Science*, **23** (3), 404-419.

30. Rubin, D. B. (2008) : For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, **2** (3), 808-840.

31. Tanner, M.A. (1993) : *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, Second Edition, New York: Springer.

32. Saville, D.J. and Wood, G.R. (1991): *Statistical Methods: The Geometric Approach*, New York: Springer.

33. Shah, K. R and Sinha, B. K. (1989): *Theory of Optimal Designs*, New York: Springer-Verlag.

34. Yang, J., Mandal, A., and Majumdar, D. (2012) : Optimal designs for two-level factorial experiments with binary response. *Statistica Sinica*, **22** (2), 885-907.

35. Yang, J. and Mandal, A. (2015) : D-optimal factorial designs under generalized linear models. *Communications in Statistics - Simulation and Computation*, **44** (9), 2264-2277.

36. Yang, J., Mandal, A., and Majumdar, D. (2016) : Optimal designs for 2^K factorial experiments with binary response. *Statistica Sinica*, **26** (1), 385-411.

37. Wu, C.F.J. (1985): Efficient sequential designs with binary data. *Journal of American Statistical Association*, **80**, 974-984.