ARTICLE IN PRESS

Journal of Statistical Planning and Inference ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

A method for augmenting supersaturated designs

Qiao-Zhen Zhang [a], Hong-Sheng Dai [b], Min-Qian Liu [a,*], Ya Wang [c]

[a] School of Statistics and Data Science & LPMC, Nankai University, Tianjin 300071, China
[b] Department of Mathematical Sciences, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK
[c] State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System (CEMEE), Luoyang, 471003, China

ARTICLE INFO

Article history:
Received 31 August 2016
Received in revised form 25 June 2018
Accepted 26 June 2018
Available online xxxx

Keywords:
Bayesian *D*-optimality
Coordinate-exchange algorithm
Follow-up experiment
Sequential design
Supersaturated design

ABSTRACT

Initial screening experiments often leave some problems unresolved, adding follow-up runs is needed to clarify the initial results. In this paper, a technique is developed to add additional experimental runs to an initial supersaturated design. The added runs are generated with respect to the Bayesian D_s-optimality criterion and the procedure can incorporate the model information from the initial design. After analysis of the initial experiment with several methods, factors are classified into three groups: primary, secondary, and potential according to the times that they have been identified. The focus is on those secondary factors since they have been identified several times but not so many that experimenters are sure that they are active, the proposed Bayesian D_s-optimal augmented design would minimize the error variances of the parameter estimators of secondary factors. In addition, a blocking factor will be involved to describe the mean shift between two stages. Simulation results show that the method performs very well in certain settings.

1. Introduction

Screening is the first phase of an experimental study on systems and simulation models. Its purpose is to eliminate negligible factors so that efforts may be concentrated upon just the important ones (active factors). Using a supersaturated design (SSD) whose run size is not enough for estimating all the main effects may be considered when a large experiment is infeasible in practice. SSDs were introduced by Box (1959), but not studied further until the appearance of the work by Lin (1993) and Wu (1993). Many developments in the area have taken place over the last two decades. For further details, please refer to Georgiou (2014), Sun et al. (2011) and the references therein.

The analysis of SSDs is challenging due to the inherent non-full rank nature of the design matrix and the fact that the columns of the model matrix are correlated. As a result, the effects of different factors are aliased with one another making it very difficult to identify the active factors correctly. Methods to overcome these problems include regression procedures, such as forward selection (Westfall et al., 1998), stepwise and all-subsets regression (Abraham et al., 1999), partial least squares methods (Zhang et al., 2007; Yin et al., 2013), shrinkage methods, including SCAD (Li and Lin, 2002) and Dantzig selector (Phoa et al., 2009) and Bayesian methods (Beattie et al., 2002; Chen et al., 2011, 2013; Huang et al., 2014). Readers can refer to Salawu et al. (2015) and Georgiou (2014). However, different methods may give different results and no method is infallible.

If we want to clarify or confirm initial results and guide the next phase of experimentation, adding follow-up runs to the initial design is a useful way. As a matter of fact, performing extra experimental runs is the only data-driven way

* Corresponding author.
E-mail address: mqliu@nankai.edu.cn (M.-Q. Liu).

https://doi.org/10.1016/j.jspi.2018.06.006
0378-3758/

Please cite this article in press as: Zhang Q.-Z., et al., A method for augmenting supersaturated designs. J. Statist. Plann. Inference (2018), https://doi.org/10.1016/j.jspi.2018.06.006.

to break confounding patterns and to disentangle confounded effects. Suppose an SSD of n_1 runs and k 2-level factors, denoted by SSD(n_1, k), has been run and now the experimenter can afford n_2 more runs to resolve ambiguities, the target is to find the best way to augment the original design to reduce uncertainty and get the most information out of the final SSD(n_1+n_2, k). Gupta et al. (2010) considered the problem for 2-level SSDs firstly: $E(s^2)$-optimal designs (proposed by Booth and Cox, 1962) are augmented with additional runs to create a new class of "extended $E(s^2)$-optimal" designs. Then Gupta et al. (2012) extended the method to s-level designs. Suen and Das (2010) also used a similar approach to add or remove one row from an existing $E(s^2)$-optimal design to make a new $E(s^2)$-optimal design. Qin et al. (2015) studied the optimality of the extended design generated by adding few runs to an existing $E(\chi^2)$-optimal mixed-level SSD and their paper covers the work of Gupta et al. (2010, 2012) as two special cases. All of these methods, however, did not consider using the information from the initial analysis and design when adding runs. Gutman et al. (2014) proposed an SSD augmentation strategy using the Bayesian D-optimality criterion, they considered the information gained from the initial design, SSD(n_1, k), as a prior, and constructed the final SSD(n_1+n_2, k) to reduce the error variances of the parameter estimators under the Bayesian paradigm.

When adding runs to fractional factorial designs, two optimality criteria, D-optimality and D_s-optimality, are often used. The D_s-optimal design approach would be applied if the experimenters emphasize precise estimation for the "subset" of the experimental factors. Kiefer and Wolfowitz (1961) defined a design as D_s-optimal if it minimizes the determinant of the normalized covariance sub-matrix of estimators of the chosen model parameters while treating the other parameters as nuisance parameters. The use of D_s-optimality designs would result in increased power since the parameters of interest are estimated more precisely (Atkinson and Donev, 1992; Casey et al., 2005). In this paper, we will combine the D_s-optimality criterion with the Bayesian technique to propose an alternative approach, which is different from the Bayesian D-optimal augmentation in two aspects: the principle of factor classification and the optimal criterion.

The next section reviews the relevant background firstly, then we propose the new algorithmic augmentation strategy for SSDs using information from the initial runs in Section 3. Section 4 compares the performance of the Bayesian D_s-optimal augmented designs with the Bayesian D-optimal augmented designs by several highlighting examples. Some concluding remarks are provided in Section 5.

2. Bayesian *D*-optimality and model selection methods

In this section, we briefly review the approach for developing augmenting Bayesian D-optimal designs in the context of linear models and some model selection methods applied to SSDs.

2.1. Bayesian D-optimality

Consider the linear model

$$\boldsymbol{y} = \beta_0\mathbf{1}_n + \beta_1\boldsymbol{x}_1 + \cdots + \beta_k\boldsymbol{x}_k + \boldsymbol{\varepsilon} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of observations, β_0 is the intercept term, $\mathbf{1}_n$ is an $n \times 1$ column vector with all elements unity, $\boldsymbol{x}_i$ is an $n \times 1$ vector of settings for the ith factor, $\boldsymbol{X}$ is the $n \times p$ design matrix with $p = k + 1$, $\boldsymbol{\beta}$ is the $p \times 1$ vector of coefficients to be estimated, and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}_n, \sigma^2\boldsymbol{I}_n)$ is the noise vector, where $\mathbf{0}_n$ is an $n \times 1$ column vector with all elements zero, and $\boldsymbol{I}_n$ is an identity matrix of order n. In a two-level factorial design, each factor setting can be coded as ±1 (or simply $\pm$). Let the prior distribution of the parameters be $\boldsymbol{\beta} \mid \sigma^2 \sim N(\boldsymbol{\beta}_0, \sigma^2\boldsymbol{R}^{-1})$, where $\boldsymbol{\beta}_0$ is the mean of prior distribution for $\boldsymbol{\beta}$, $\boldsymbol{R}$ is a prior covariance matrix, and the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{\beta}$ be $\boldsymbol{y} \mid (\boldsymbol{\beta}, \sigma^2) \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$. Then the posterior distribution for $\boldsymbol{\beta}$ given $\boldsymbol{y}$ is

$$\boldsymbol{\beta} \mid \boldsymbol{y} \sim N(\boldsymbol{b}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{R})^{-1}),$$

where $\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{R})^{-1}(\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{R}\boldsymbol{\beta}_0)$.

Let $\boldsymbol{X}_1$ be a model matrix corresponding to the initial n_1 runs of an experiment with response vector $\boldsymbol{y}_1$, and $\boldsymbol{X}_2$ be the additional n_2 rows with response vector $\boldsymbol{y}_2$. That is

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}, \ \boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix}.$$

Once the data from the first stage have been collected, many different analysis methods can be employed to identify active factors and the information from the analysis may be used as a prior. Gutman et al. (2014) pointed out that the experimenter can classify a factor as primary term (highlighted by an analysis method or many methods), secondary term (if there is an indication the factor may be active, but it is not a predominant), or potential term (with little evidence to suggest it is active). Then prior distributions would be assigned as follows. Since the primary terms are likely to be active, their coefficients are specified to have a diffuse prior variance tending to infinity (DuMouchel and Jones, 1994), which implies that the primary terms are likely to be much different from zero. On the other hand, potential terms are unlikely to have large effects, and it is proper to assume that they have a relative small variance. For secondary terms, they may or may not be active, so their prior variances should be finite, but larger than that for potential terms. We assume that the factors in $\boldsymbol{X}$ have been reordered after

Please cite this article in press as: Zhang Q.-Z., et al., A method for augmenting supersaturated designs. J. Statist. Plann. Inference (2018), https://doi.org/10.1016/j.jspi.2018.06.006.