OVERVIEW OF OPTIMAL EXPERIMENTAL DESIGN AND A SURVEY OF ITS EXPANSE IN APPLICATION TO AGRICULTURAL STUDIES

Stephen J. Walsh
Department of Mathematics and Statistics, Utah Stat University, Logan, UT
steve.walsh@usu.edu
www.walsh-datascience.com

ABSTRACT

Optimal Design of Experiments is currently recognized as the modern dominant approach to planning experiments in industrial engineering and manufacturing applications. This approach to design has gained traction among practitioners in the last two decades on two-fronts: 1) optimal designs are the result of a complicated optimization calculation and recent advances in both computing efficiency and algorithms have enabled this approach in real time for practitioners, and 2) such designs are now popular because they allow the researcher to 'design for the experiment' by working constraints, cost, number of experiments, and the model of the intended post-hoc data analysis into the design definition, thereby creating designs with more practical meaning than classical or catalogue designs. In this talk, I will review the definition of optimal design, discuss recent computational advancements in this field, and provide a survey of the expanse of this design approach in the agricultural literature.

KEYWORDS: Optimal Design of Experiments, Classical Experimental Design, D-criterion, I-criterion

1 Introduction

Development of the practice of statistical design of experiments (DoE) is historically influenced by three broad applied science domains: 1) Agriculture, 2) Industrial Engineering and Manufacturing, and 3) Bio-med (pharmaceutical development, medicine, and epidemiology). A recent paper by Jensen (2018) in the journal *Quality Engineering* summarized the last several decades of developments in DoE research, approach, and application. In the last two decades Jensen highlights a movement of industrial practitioners away from *classical designs* (e.g. 'catalogue designs' which include the factorial family, central-composite designs, and standard definitions of block, split-plot designs and others) toward *optimal designs*.

Classical designs trace back to Sir R. A. Fishers foundational work *The Design of Experiments* (Fisher 1935). This book formalized the analysis-of-variance, discusses the importance of the homogeneity of variance assumption, develops designs for blocking, the Latin square, the factorial family, addresses confounding, and provides applications of data collection and analysis. The practice and teaching pedagogy of DoE is still largely consistent with the structure of this book at most universities.

The conceptualization and definition of optimal design predates *The Design of Experiments*. In 1918 Kirstine Smith, a Danish statistician working for Karl Pearson, published what is now recognized as the seminal paper for the optimal design concept and was 30 years ahead of its time. In this paper Smith discussed how viewing the residual error-variance as a function of the design levels of the experimental variables informs 'minimum uncertainty' designs for a single-factor experiment and from the perspective of fitting a first- up to a sixth-order polynomial to the experimental data. This approach to design did not gain widespread attention in the following decades because the optimal design is found as the optimization of a complicated non-convex multidimensional objective function. Not until recent decades were computing power and appropriate optimization routines available to firmly put optimal design in the grasp of the experimental practitioner.

In the remainder of this paper, we will provide a definition of optimal DoE, discuss its attractiveness to practitioners in contrast to classical designs, and provide a survey of several agricultural journals and investigation into the expanse of application of optimal DoE in agricultural studies.

2 Framing the Optimal Design Problem

The attractiveness of the optimal DoE perspective is that it allows the practitioner to 'design for the experiment' by considering specific/unique aspects of the current experiment as opposed to 'experimenting for the design' by ensuring that implemented runs conform to a catalogue design (Goos and Jones 2011). In this respect the following specific aspects of the experiment are considered in the design formulation and constitute inputs to the design generation:

1. N: number of affordable experimental runs,
2. the model with which the resulting experimental data will be analyzed, and
3. a definition of what a 'good' design is for your specific study.

These three quantities are taken as inputs to on optimization calculation which attempts to give a design that minimized the portion of uncertainty in the study that is attributable to the specific sample locations of the study factors. To formalize this, we will illustrate the concept from the matrix-algebra framework of standard linear models.

Let K:= number of experimental factors. Then a design is the $N \times K$ design matrix $\mathbf{X}$. The objective of the optimal design calculation is to populate the rows of $\mathbf{X}$ thereby giving the set of experiments that should be implemented in the study. A linear model with standard assumptions is assumed and will be used to analyzed the experimental data:

$$\mathbf{y} = \mathbf{F}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{y}$ is the $N \times 1$ vector of responses, $\mathbf{F}$ is the $N \times p$ *model matrix* and is an expansion of design matrix $\mathbf{X}$, $\boldsymbol{\beta}$ is the $p \times 1$ parameter vector that will be estimated from the data, and $\boldsymbol{\varepsilon} \sim \mathcal{N}_p(\mathbf{0}, \sigma^2\mathbf{I})$ where σ^2 represents the residual error variance.

Constructing the optimal design criterion requires a solid understanding of the relationship between design matrix $\mathbf{X}$ and model matrix $\mathbf{F}$. The model matrix is a function of the design matrix and so some authors write $\mathbf{F}(\mathbf{X})$ which represents an expansion of the rows of $\mathbf{X}$ into the model form encoded by $\mathbf{F}$. To be precise, consider the situation where the researcher is experimenting on $K = 2$ factors and they intend to fit a second order polynomial model to the experimental data. The scalar linear model can be written, for observation $i = 1, \dots, N$

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{1i}x_{2i}\beta_{12} + x_{1i}^2\beta_{11} + x_{2i}^2\beta_{22} + \varepsilon_i. \tag{2}$$

In this scenario the ith row of the design matrix $\mathbf{X}$ is

$$\mathbf{x}_i' = (x_{1i} \quad x_{2i}) \tag{3}$$

and the corresponding row of the model matrix is $\mathbf{F}(\mathbf{X})$

$$\mathbf{f}_i' = \left(1 \quad \mathbf{x}_i \quad x_{1i}^2 \quad x_{2i}^2\right). \tag{4}$$

Under this model framework, Fisher's Information Matrix for this model is

$$\mathbf{M}(\mathbf{X}) = \mathbf{F}'\mathbf{F} \tag{5}$$

and an optimal design is that $\mathbf{X}$ which optimizes some function of $\mathbf{M}$ thereby 'maximizing the information gained from the experiment. Eq. (5) makes explicit that the model information is solely a function of the implemented experimental design $\mathbf{X}$. The importance of the design via the information matrix is easily seen in the regression estimating equations, where

$$\widehat{\boldsymbol{\beta}} = (\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'\mathbf{y} \tag{6}$$
$$\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right) = \sigma^2(\mathbf{F}'\mathbf{F})^{-1} \tag{7}$$

are the formula for the regression coefficient estimate and its variance-covariance matrix respectively. Further, if the practitioner intends to use the model fitted to the experimental data for prediction, the variance of the mean predicted value $\hat{\mathbf{y}}$ at a new design point $\mathbf{x}_{\mathrm{new}}'$ is

$$\mathrm{Var}\left(\hat{\mathbf{y}}\ (\mathbf{x}_{\mathrm{new}}')\right) = \sigma^2\mathbf{f}'(\mathbf{x}_{\mathrm{new}}')(\mathbf{F}'\mathbf{F})^{-1}\mathbf{f}(\mathbf{x}_{\mathrm{new}}') \tag{8}$$

which illustrates that the design propagates into the prediction uncertainty because it is embedded in the model matrix, i.e. $\mathbf{F}(\mathbf{X})$. Eq.s (7) and (8) show that the uncertainty quantification of regression coefficients and mean predictions are a function of two quantities: 1) the residual error variance σ^2 which represents the total combined uncertainty of the measurement device with variance attributable to experimental repetition, and 2) the information matrix $\mathbf{M}(\mathbf{X}) = \mathbf{F}'\mathbf{F}$ which is a direct result of the implemented experimental design $\mathbf{X}$. This perspective makes clear that it is entirely possible that the experimental design chosen by the practitioner, that is, the specific settings of the experimental factors, may in fact be the dominant source of uncertainty in the experimental data! Thus, it behooves the experimenter to choose the specific factors levels carefully. This idea of 'careful choice of experimental levels', that is, the treatments, is the essence of the optimal design perspective.

2.1 Common Optimal Design Criteria

The designer must specify an optimality criterion which encodes the meaning of a 'good design' for their specific purpose. There are many ways to approach this problem, and we here briefly describe two of the most common implemented design criterion, the D- and I-criterion.

The D-criterion is used to find an experiment which will give high precision on the regression coefficients $\widehat{\boldsymbol{\beta}}$, or, in a sense, the smallest $\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{F}'\mathbf{F})^{-1}$. The D-criterion gets its name as it is the determinant of the inverse of the information matrix, or

$$D(\mathbf{X}) = 1/|\mathbf{F}'\mathbf{F}| \tag{9}$$

and a D-optimal design, denoted $\mathbf{X}^*$is defined as

$$\mathbf{X}^* = \arg\min_{\mathbf{X}} D(\mathbf{X}). \tag{10}$$

If, instead, the primary use of the fitted regression model is to make post-experiment predictions, then it is common for practitioners to use I-optimal designs. The I-criterion is a candidate designs average scaled prediction variance, formally

$$I(\mathbf{X}) = \frac{N}{V}\int \mathbf{f}'(\mathbf{x}')(\mathbf{F}'\mathbf{F})^{-1}\mathbf{f}(\mathbf{x}')\,\mathbf{dx}' \tag{11}$$

where V is the volume of the design space. Thus, an I-optimal design is defined as

$$\mathbf{X}^* = \arg\min_{\mathbf{X}} I(\mathbf{X}). \tag{12}$$

2.2 Algorithms for Generating Optimal Designs

Optimization searches for globally optimal designs via objective functions like the D- and I-criterion as stated in Eq.s (11-12), are well-known as difficult high-dimensional non-convex optimality searches. One of the reasons for the recent widespread adoption of the optimal design approach in industrial engineering and manufacturing is the availability of fast robust algorithms for searching these criteria for the optimal design. In this regard, the coordinate exchange algorithm, appears to be the most successful and widely applied (Goos and Jones 2011). More recently, Walsh (2021), has produced results indicating the particle swarm optimization may be superior to existing algorithms for finding globally optimal designs (Walsh 2021, Walsh and Borkowski 2022, Walsh and Borkowski 2022).

3 Survey of the Agricultural Peer-Reviewed Literature

The purpose of this study was to investigate the statistics literature in the agricultural sciences to assess the degree to which the optimal design perspective has been adopted and applied. We selected the following agricultural statistics and application journals to review:

1. Journal of Agricultural and Biological Statistics,
2. Journal of Animal Science,
3. Plant Biotechnology Journal,
4. Field Crops Research,
5. Plant Science, and
6. Journal of Dairy Science.

We used each journals webpage to search the following keywords: *optimal design of experiments, optimal experimental design, modern experimental design, D-criterion*. If in the particular journal these keywords did not yield many searches, we searched typical classical design of experiments nomenclature to ensure that these journals indeed had a high volume of experimental design publications. In that regard we searched keywords: *Response Surface Methodology (RSM), factorial design, randomized block design, split-plot, ANOVA, and repeated measures*.

The most successful search was in the Journal of Agricultural and Biological Statistics (JABES). (Coffey and Gennings 2007) provide a paper discussing the use of D-optimal designs for mixed and continuous outcomes analyzed via non-linear models. The application of this paper focused on dose-response applications. (Parker and Gennings 2008) provide an extension to generating optimal designs for dose-response studies using penalized local optimal designs. (Zolghadr and Zuyev 2016) provide a paper discussing the use of Bayesian optimal design for dilution experiments under volume constraints. (Shotwell and Gray 2016) discuss the optimal design perspective in dynamic multi-scale model applications for studying cardio electrophysiology. A special issue in JABES 2020 focused on design experiments in agriculture. Only two papers in this issue discussed optimal design. (Huang, et al. 2020) discuss applications of optimal design for non-linear models using the Michaelis-Menten Kinetics phenomenology. Last, an excellent survey of the history of experimental design in agriculture is given by (Verdooren 2020). However, this paper does not highlight any widespread or single use (for that matter) of optimal design. Simply, the last sentence of the paper makes a mention to the optimal design concept and references a 15 year old text by Atkinson on the topic.

Searches in the Journal of Animal Science yielded a few hits on optimal design searches. All three papers found discussed the application of the optimal design concept to animal breeding experiments (Olson, Willham and Boehlje 1980, Sölkner 1993, Lozano-Jaramillo, et al. 2020).

Searches for the optimal design concept in the remaining journals, Plant Biotechnology Journal, Field Crops Research, Plant Science, and Journal of Dairy Science, yielded no hits. We did confirm via searches for classical design nomenclature that these journals do in fact contain a large density of publications on design of experiments and applications.

4 Opportunities for Optimal Design of Experiments in Agricultural Applications

Given the popularity of the optimal design concept in industrial settings in conjunction with the fact that agricultural experimentation has probably generated more statistical research on the topic, we were somewhat surprised that our literature survey did not turn up more results on applications of optimal design in agricultural studies. The reason is unknown, and we would be interested to discuss this with you further. Note that several professionals did approach us at AgStats 2022 conference indicating that further searches in the animal breeding literature would elucidate a more widespread application of the optimal design concept in that field.

5 Discussion and Conclusions

In this paper we highlighted that one of the major experiment application domains, industrial engineering and manufacturing, has markedly moved toward the optimal design concept over classical designs due in part to the availability of cheap computing and robust algorithms for solving the optimal design problem. We presented the basic definition and formulation of optimal design, and illustrated its perspective via the typical linear-regression framework. In short, optimal designs are those that minimize the portion of uncertainty in the data analysis that is attributable directly to the actual experiment that the practitioner has chosen to implement. Thus, an optimal design is a prudent choice because the experiment itself is one of two main components of uncertainty in the analysis. Optimal designs make the most sense if at least one of the experimental factors is on a continuous scale, and the result of the optimal design calculation is the precise best locations to set the treatments of such a factor. It is possible to produce optimal designs for categorical factors, e.g. such as finding the optimal allocation of replication in the different levels of a split-plot structure. Given the results of this literature survey, it appears that there is a large opportunity to explore further the application of the optimal design concept in agricultural studies.

References

Coffey, T. , and D. Gennings. 2007. "D-optimal designs for mixed discrete and continuous outcomes analyzed using nonlinear models." *Journal of Agricultural, Biological, and Environmental Statistics.*

Fisher, R. A. 1935. *The Design of Experiments.* Macmillan.

Goos, P., and B. Jones. 2011. *Optimal Design of Experiments: A Case Study Approach.* West Sussex, United Kingdom: John Wiley and Sons, Ltd.

Huang, Yuanzhi, S. G. Gilmour, K. Mylona, and P. Goos. 2020. "Optimal Design of Experiments for Hybrid Nonlinear Models, with Applications to Extended Michaelis–Menten Kinetics." *Journal of Agricultural, Biological and Environmental Statistics* 601-616.

Jensen, W. A. 2018. "Open problems in optimal design." *Quality Engineering* 583-593.

Lozano-Jaramillo, Maria, Hans Komen, Y. C. J. Wientjes, H. A. Mulder, and J. W. Bastiannsen. 2020. "Optimizing design to estimate genetic correlations between environments with common environmental effects." *Journal of Animal Science.*

Olson, T. A., R. L. Willham, and M. D. Boehlje. 1980. "Linear programming as a tool for planning beef cattle breeding experimentation." *Journal of Animal Science* 847-859.

Parker, S. M., and C. Gennings. 2008. "Penalized Locally Optimal Experimental Designs for Nonlinear Models." *Journal of Agricultural, Biological, and Environmental Statistics* 334-354.

Shotwell, M. S., and R. A. Gray. 2016. "Estimability Analysis and Optimal Design in Dynamic Multi-scale Models of Cardiac Electrophysiology." *Journal of Agricultural, Biological, and Environmental Statistics* 1-16.

Smith, K. 1918. "ON THE STANDARD DEVIATIONS OF ADJUSTED AND INTERPOLATED VALUES OF AN OBSERVED POLYNOMIAL FUNCTION AND ITS CONSTANTS AND THE GUIDANCE THEY GIVE TOWARDS A PROPER CHOICE OF THE DISTRIBUTION OF OBSERVATIONS." *Biometrika* 1-85.

Sölkner, Johann. 1993. "Choice of optimality criteria for the design of crossbreeding experiments." *Journal of Animal Science* 2867-2873.

Verdooren, L. R. 2020. "History of the Statistical Design of Agricultural Experiments." *Journal of Agricultural, Biological and Environmental Statistics* 457-486.

Walsh, Stephen J. 2021. *Development and applications of particle swarm optimization for constructing optimal experimental designs.* Ph.D. Dissertation, Bozeman, MT: Montana State University.

Walsh, Stephen J., and John J. Borkowski. 2022. "Fast Computation of Highly G-optimal Exact Designs via Particle Swarm Optimization." *arXiv preprint arXiv:2206.06498.*

Walsh, Stephen J., and John J. Borkowski. 2022. "Generating Exact Optimal Designs via Particle Swarm Optimization: Assessing Efficacy and Efficiency via Case Study." *arXiv preprint arXiv:2206.06940.*

Zolghadr, M., and Sergei Zuyev. 2016. "Optimal Design of Dilution Experiments Under Volume Constraints." *Journal of Agricultural, Biological and Environmental Statistics* 663-683.