Focus Article

Experimental design

J. P. Morgan* and Xinwei Deng

Maximizing data information requires careful selection, termed *design*, of the points at which data are observed. Experimental design is reviewed here for broad classes of data collection and analysis problems, including: fractioning techniques based on orthogonal arrays, Latin hypercube designs and their variants for computer experimentation, efficient design for data mining and machine learning applications, and sequential design for active learning. © 2012 Wiley Periodicals, Inc.

How to cite this article:
WIREs Data Mining Knowl Discov 2012, 2: 164–172 doi: 10.1002/widm.1046

DESIGN FOR EFFICIENT KNOWLEDGE DISCOVERY

Experimental design is the subfield of statistics concerned with information optimization in scientific investigations. Design activities frequently take place prior to data collection in the following broadly described framework. There is a response variable y of interest, and there are controllable variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$ conjectured to affect that response. The design problem is to select values of $\boldsymbol{x}$, say $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ for $i = 1, \ldots, N$, at which to observe values of y. This selection is guided by maximizing one or more measures of information that will be gained on y and its relationship with the x_js. Design techniques for data mining applications, in which large amounts of data $(y_i, \boldsymbol{x}_i)$ could be, will, or have already been, collected, have the goal of more efficient exploration of relationships in that data. Included are *restricted data mining situations* in which a vast amount of data could potentially be gathered, but cost or other considerations demand a much more frugal approach. In restricted situations, design strategies can foster keen insight into the $y, \boldsymbol{x}$ relationship while collecting a very small fraction of the possible data.

A useful construct is that of a *factorial experiment*, or experiment with factorial treatment structure, which is just an experiment with p controllable variables as described above. Let S_j be the collection of possible values for factor x_j and, for simplicity, assume that all members of $\mathcal{S} = S_1 \times S_2 \times \cdots \times S_p$ are candidates for $\boldsymbol{x}$. The design problem of choosing $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ at which to measure y has the goal of understanding (some features of) the relationship $y = f(\boldsymbol{x}) + \epsilon$, ϵ representing measurement noise. Each S_j may be finite, countable, or uncountably infinite. Regardless, the number of possible data locations $|\mathcal{S}|$ can be unmanageably large. The design problem in a restricted data mining situation is to 'mine' the data field $\mathcal{S}$ in such a way that a faithful representation of $f(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{S}$ can be created.

Design's historical roots stretch back to the birth of statistics, with many fundamental concepts originating with Fisher.[1] Rigorous study of optimal design selection grew rapidly from the seminal works of Elfving,[2] Kiefer,[3] Kiefer and Wolfowitz,[4] and Box and Draper,[5] among others. In brief, if the functional form of f is known up to parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$, then information assessment can be based on the inverse of the covariance matrix for estimators $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, commonly called the *information matrix*, denoted as $I(\boldsymbol{\beta})$. Assuming $I(\boldsymbol{\beta})$ has the same, maximal rank k for all competing designs (designs yielding smaller rank are eliminated), denote its nonzero eigenvalues by $\lambda_1, \lambda_2, \ldots, \lambda_k$. Popular optimality measures for information content, all of which are direct measures of variance magnitude and so are to be minimized, are $A = \sum_{i=1}^{k} \lambda_i^{-1}$, $D = \prod_{i=1}^{k} \lambda_i^{-1}$, and $E = \max_i \lambda_i^{-1}$. These are proportional respectively to the average variance of the $\hat{\beta}_i$, volume of the confidence ellipsoid for $\hat{\boldsymbol{\beta}}$, and maximal variance over all normalized linear combinations of the $\hat{\beta}_i$. Variants on these and many other criteria are discussed by, for example, Atkinson et al.,[6] Morgan and Wang,[7] and Gilmour and Trinca.[8] If the postulated form of f is incorrect then $\widehat{\boldsymbol{\beta}}$ will be biased for $\boldsymbol{\beta}$ and design considerations can incorporate notions of bias abatement, taken up for fractions in the following.

*Correspondence to: jpmorgan@vt.edu
Department of Statistics, Virginia Tech, Blacksburg, VA, USA
DOI: 10.1002/widm.1046

FINITE DATA FIELDS AND FRACTIONAL FACTORIAL DESIGNS

If each S_j is finite, $|S_j| = s_j$, then we have a $|\mathcal{S}| = s_1 \times s_2 \times \cdots \times s_p$ factorial. In many arms of scientific endeavor, it is not uncommon to replace any interval S_j with an appropriately spaced subset of its points to arrive at this situation. At the boundary of this approach is the commonly employed 2^p framework in which each factor is considered at $s_j = 2$ levels, these covering, or nearly covering, the actual ranges of the factors. Use of two-level factors minimizes $|\mathcal{S}|$ for given p, yet even then the cost of data collection may preclude measuring at every point in $\mathcal{S}$. Fractioning is a technique for dealing with this restricted situation.

For finite $\mathcal{S}$, the values $f(\boldsymbol{x})$ relating y to x may be thought of as a finite collection of arbitrary means $f(x_1, x_2, \ldots, x_p) = \mu_{x_1 x_2 \ldots x_p}$. Indexing the values of x_j by 0, 1,...,$s_j - 1$, these means may be collectively written as the vector $\boldsymbol{\mu} = (\mu_{00\ldots0}, \mu_{00\ldots1}, \ldots, \mu_{s_1-1,s_2-1,\ldots,s_p-1})$. Fractioning takes advantage of the fact that $\boldsymbol{\mu}$ may admit a lower-dimensional representation. Let $\mathcal{J}$ be the set of 2^p subsets of the indices $\{j_1, j_2, \ldots, j_p\}$. For $J \in \mathcal{J}$ there are $\Pi_{j \in J} s_j$ marginal means defined as the average of $\mu_{j_1 j_2 \cdots j_p}$ over all values of the $p - |J|$ subscripts not in J; there is one such mean if $|J| = 0$, and otherwise one such mean for each distinct set of values for j_js in J. A *main effects model* represents $\mu_{j_1 j_2 \cdots j_p}$ as a linear combination of the marginal means for which $|J| \leq 1$. For $q \geq 2$, a *q-factor interaction model* represents $\mu_{j_1 j_2 \cdots j_p}$ as a linear combination of the marginal means for which $|J| \leq q$. Explicit expressions based on orthogonal parameterizations may be found in Dey and Mukerjee[9] or Hedayat et al.[10] A p-factor interaction model is equivalent to arbitrary $\boldsymbol{\mu}$.

A q-factor interaction model for $q \leq p$ can be unbiasedly fit without mining all points in $\mathcal{S}$, provided the correct points are selected. For simplicity, we proceed with $s_j = s$ for all j, and $|\mathcal{S}| = s^p$. An orthogonal array (OA) of strength t, OA(N, s^p, t), is a collection of N points in $\mathcal{S}$ which, when written as the rows of an $N \times p$ matrix, contains as the rows of each $N \times t$ submatrix each of the s^t combinations of the corresponding t factors with frequency N/s^t (see Figure 1). Observation at the points (rows) of an OA of strength t will unbiasedly estimate any $f(\boldsymbol{x})$ that can be expressed as a u-factor interaction model for some $u \leq \lfloor t/2 \rfloor$. Moreover, this estimation is optimal with respect to the A, D, and E criteria and many others.[11,12] The orthogonality property implies that model reduction via removal of any subset of the terms in the u-factor model will not affect the estimates for the remaining terms. Orthogonal arrays are

$$\begin{matrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{matrix} \qquad \begin{matrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 2 \\ 0 & 2 & 2 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 2 & 0 \\ 1 & 2 & 0 & 2 \\ 2 & 0 & 2 & 2 \\ 2 & 1 & 0 & 1 \\ 2 & 2 & 1 & 0 \end{matrix}$$

$\mathrm{OA}(8, 2^4, 3)$ $\mathrm{OA}(9, 3^4, 2)$

FIGURE 1 | Two orthogonal arrays.

by a wide measure the most commonly employed of the *fractional factorial designs*.

OAs have been extensively studied for several decades (see the book-length treatment,[10] the online orthogonal array catalogs,[13,14] and the recent survey[15]). Use of OA fractions of strength t when $u > \lfloor t/2 \rfloor$ will introduce bias which can be minimized in any of several reasonable senses by choice of particular OAs; especially popular have been various versions of the *minimum aberration* criterion.[16–19] Interestingly, blended criteria that incorporate measures of both variance and bias can lead to designs other than OAs as best.[20] In any case, the real value of fractioning via OAs rests on the empirical fact that many data situations do not demand a fully general model, rather they are amenable to a lower-order interaction model, to gain an adequate representation. Even with $u > \lfloor t/2 \rfloor$ biases may be sufficiently mild that the essential behavior of $f(\boldsymbol{x})$ is not obscured.

When considering many factors (large p), *screening designs* employ the main effects model with only $s = 2$ levels per factor to economically determine which factors exert strong (main) effects. The two-level OAs of strength 2, OA(N, 2^p, 2), must have N a multiple of 4 and can accommodate up to $p = N - 1$ factors. These arrays have been fully enumerated for up to $N = 20$ runs,[21] for $N = 24$ and $p \leq 7$, for $N = 28$ and $p \leq 6$, and for $N = 32$ and $p \leq 6$.[22] Full enumeration enables design selection through exhaustive comparisons of bias measures and projection properties. Depending on the criterion selected, the best OAs for a range of larger values of N and p have been determined without full enumeration.[23–26]

Supersaturated designs have $p \geq N$ and hence necessarily risk bias even should a main effects model hold. They have nonetheless proven to be a valuable experimental tool when relatively few of the factors are expected to exert significant influence on the response y. Many techniques for devising effective supersaturated designs, including many

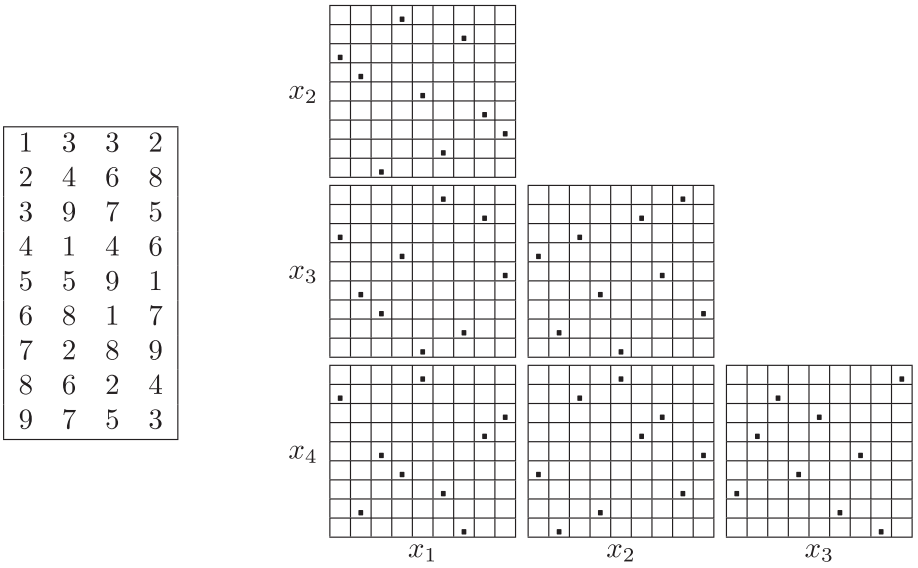

FIGURE 2 | A four-dimensional, nine-level Latin hypercube and its two-dimensional projections.

based on modifications of two-level OAs, have been devised.[27–34]

COMPUTER EXPERIMENTS

Highly complex physical phenomena on many scales, from weather systems to artificial limb function, have increasingly become subjects of investigation through computer modeling. Elaborate, deterministic mathematical models are proposed and coded for computation, usually involving numerous differential and/or integral equations and boundary conditions, and depending on numerous inputs $\boldsymbol{x} = (x_1, \ldots, x_p)$. A single computer run for a model with a single input $\boldsymbol{x}$ may be quite time-consuming, and regardless, the design space $\mathcal{S}$ is usually infinite. Thus from the potentially infinite set of observations $(y, \boldsymbol{x})$, and similar to the fractioning problem explored above, we are faced with the restricted problem of selecting the points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ at which runs will actually be made. But the truly distinguishing feature of a computer experiment is the lack of error: all runs at the same input $\boldsymbol{x}$ will produce the same output y. Consequently, replication of inputs is to be avoided, even on subsets of the input variables $\boldsymbol{x}$, for should the relationship $y = f(\boldsymbol{x})$ not involve some of the predictors, replication on the remaining subset provides no information. Both modeling and design issues for computer experiments are taken up at length by Santner et al.[35] and Fang et al.[36] Selection of N is discussed by Loeppky et al.[37] Popular design strategies will be reviewed here.

McKay et al.[38] introduced *Latin hypercube designs* (LHDs) for computer experimentation. The range of each x_j is divided into N intervals of equal length. These intervals are numbered by their midpoints, randomly ordered, then assigned as columns of the $N \times p$ design matrix $\boldsymbol{X}$. Rows of $\boldsymbol{X}$ are the N inputs $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ of the LHD (see Figure 2).

LHDs are simple to construct and have the desired uniformly distributed, nonreplicated, one-dimensional projections. These two attributes have made them very popular, and several modifications to impose additional, desirable properties on the basic design have followed. A *space-filling* or *maximin* LHD[39–42] is a LHD which, for given p and N, maximizes the minimum Euclidean distance between any two $\boldsymbol{x}_i$. The intent of space filling is to achieve an improved representation of $f(\boldsymbol{x})$ by more evenly covering the space $\mathcal{S}$. *Orthogonal LHDs* (OLHDs)[43–47] have orthogonality of the columns of $\boldsymbol{X}$, assuring orthogonal estimation of effects in a first-order polynomial model for $f(\boldsymbol{x})$. In some cases, orthogonality under higher-order polynomial models is also achieved. More flexible are the transparently named *nearly orthogonal Latin hypercube designs* (NOLHDs).[48–50] Stratification in t-dimensional margins of a Latin hypercube for $t \geq 1$ can be achieved by constructions based on OAs of strength t.[51–53] The LHD in Figure 2 is of this type, based on the second OA in Figure 1 (viewing each 9×9 margin as a 3×3 array of 3×3 subarrays, each subarray contains exactly one point). A *sliced Latin hypercube design* (SLHD) $\boldsymbol{X}$ with slices $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ is a LHD for which the whole design $\boldsymbol{X} = (\boldsymbol{X}_1', \boldsymbol{X}_2', \ldots, \boldsymbol{X}_k')'$ is an LHD, and each

slice $\boldsymbol{X}_k$ is an LHD. SLHDs can incorporate qualitative predictors.[54] LHDs have also been developed for various sequential strategies.[55,56]

Alternatively, one may directly optimize choice of $\boldsymbol{x}_i$s with respect to a space-filling criterion. Included here are sliced space-filling designs[57] and Sudoku-based space-filling designs.[58] *Uniform designs*[36,59,60] also take a space-filling approach. Thinking of the $\boldsymbol{x}_i$s as a random sample, a uniform design minimizes a measure of the distance of the empirical distribution function of this sample from that of a continuous uniform distribution in p-space. Not surprisingly, the designs obtained depend on the distance measure employed.

DATA MINING AND MACHINE LEARNING

Again, the techniques of experimental design are built around ideas of effective data collection. The value of design strategies in data mining applications is found in tying the notion of effective collection from $\mathcal{S}$ with that of efficient exploration of relationships in $(y_i, \boldsymbol{x}_i)$. Design can advance the effectiveness of machine learning methods by enhancing accuracy and reducing variability.

Design for model discrimination has been well studied.[61] Methods have been developed to discriminate among candidate models, for instance, nested polynomial regression models with different orders.[62,63] Bingham and Chipman[64] take an optimal design approach for discriminating models under a Bayesian formulation for the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is the vector of responses, $\boldsymbol{X}$ as above is the N-rowed design matrix, $\boldsymbol{\beta}$ is the parameter vector, and $\boldsymbol{\epsilon}$ is a vector of random errors. Label the possible models as $M_1, \ldots, M_K$. To evaluate a design's capability of discriminating models, they employ the distance criterion

$$HD = \sum_{i<j} P(M_i)P(M_j)H(f_i, f_j),$$

where $P(M_i)$ is probability of model M_i and $H(f_i, f_j) = \int (f_i^{1/2} - f_j^{1/2})^2 \mathrm{d}y$ is the Bhattacharyya–Hellinger distance[65–67] between the predictive densities f_i and f_j of the response y under models M_i and M_j. A design maximizing HD will more readily identify a set of active predictors. However, the number of models that can be entertained in this framework is relatively small.

With advancing technology, fields as disparate as biology and financial services are working with massive, high dimensional data, where both N and p can be very large. This calls for variable selection to identify significant x_is for the model. Regularization methods for variable selection have received considerable attention, including Lasso,[68] nonnegative garrotes,[69] SCAD,[70] LARS,[71] and the Dantzig selector[72] among many others. The great number of candidate models in these situations precludes more traditional design approaches for model discrimination discussed above. Now the design perspective provides value by pinpointing data subsets with desirable structures, helping identify significant variables more efficiently.

Deng et al.[73] consider selection of the design matrix $\boldsymbol{X}$ for the Lasso procedure. For the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, the Lasso estimates $\boldsymbol{\beta}$ by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_{l_1},$$

where $\|\boldsymbol{\beta}\|_{l_1} = \sum_{i=1}^{p} |\beta_i|$ and λ is a tuning parameter. Because the l_1 norm $\|\cdot\|_{l_1}$ is singular at the origin, some coefficients of $\hat{\boldsymbol{\beta}}$ are exactly zero, leading to simultaneous estimation and model selection. Deng et al.[73] take $\boldsymbol{X}$ to be a nearly orthogonal LHD. Owing to their orthogonality and stratification properties, the use of NOLHDs in the Lasso can significantly improve accuracy in identifying active predictors. Xing et al.[74] provide an optimal design strategy for variable selection under Lasso for two-level designs. For observational data, one seeks a well-structured subset of the data to improve Lasso variable selection.

For data where a linear model is not realistic, machine learning techniques are valuable tools. Here, too, experimental design concepts can be incorporated to improve performance. MacKay[75] and Cohn[76] applied optimal design techniques to neural networks using linear approximations of neural network models. However, a first-order approximation can be overly rough for a complicated model. Gilardi and Faraj[77] developed a *query-by-committee* method[78] to select design points for the multilayer perceptron (MLP) model[79] in a regression setting. Given a training set $\mathcal{T} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, with the goal of modeling the unknown $f(\boldsymbol{x})$, the problem is to select new data points achieving maximal modeling information. A committee of m models $\hat{f}^{(1)}, \ldots, \hat{f}^{(m)}$ is constructed using the training set $\mathcal{T}$, then new points selected which maximize disagreement among those models. Disagreement $D(\boldsymbol{x})$ is calculated using a sample variance of the estimates at $\boldsymbol{x}$:

$$D(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}^{(i)} - \bar{y})^2,$$

where $\hat{y}^{(i)} = \hat{f}^{(i)}(\boldsymbol{x})$ is the prediction at $\boldsymbol{x}$ using $\hat{f}^{(i)}$, and $\bar{y} = (1/m)\sum_{i=1}^{m} \hat{y}^{(i)}$ is the mean of the estimates from the committee. This strategy is expected to maximize information gain by selecting points that efficiently drive convergence toward a single MLP regression model.

Support vector machines (SVMs) are another valuable technique for nonlinear models in classification and regression.[80] SVMs often need to specify values of meta-parameters, which can have a profound effect on prediction performance. Experimental design principles can be used to identify optimal parameter settings more effectively than an exhaustive grid search. For example, Staelin[81] employs design concepts for selecting meta-parameter values, which is robust and works efficiently on a variety of problems.

In data mining and machine learning, *cross-validation* is widely used to assess prediction error.[82] For a loss function $L(y, \hat{f})$ measuring the discrepancy between the predicted response $\hat{f}$ and the actual response y, the prediction error is defined to be $\gamma = E\{L(y, \hat{f}(\boldsymbol{x}))\}$. The objective of cross-validation is to estimate γ based on a training sample $\mathcal{T} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$. This is done by partitioning the sample $\mathcal{T}$ into k folds, $\mathcal{C}_1, \ldots, \mathcal{C}_k$, then iteratively holding onefold $\mathcal{C}_u$ for testing the prediction accuracy of $\hat{f}$ constructed with data in the other folds. Specifically, the estimate $\hat{\gamma}$ is computed as

$$\hat{\gamma} = \frac{1}{n}\sum_{u=1}^{k}\sum_{i\in\mathcal{C}_u} L(y_i,\ \hat{f}_{\mathcal{T}_{-\mathcal{C}_u}}(\boldsymbol{x}_i)),$$

where cross-validation iteration u uses $\mathcal{T}_{-\mathcal{C}_u} = \mathcal{T}\backslash\mathcal{C}_u$ for model building and $\mathcal{C}_u$ for model testing. From a design perspective, it is helpful to embed desirable structure into the partitioned data such that the points in each fold have attractive properties. Deng and Qian[83] introduce *sliced cross-validation* (SCV) for efficiently estimating the prediction error for classification. SCV uses SLHDs $\boldsymbol{X}$ to impose a slicing structure for the inputs in $\mathcal{T}$ so that the input values $\boldsymbol{X}_k$ in each fold are well spread in the whole space. By embedding each slice $\boldsymbol{X}_k$ into a corresponding fold $\mathcal{C}_k$, SCV can reduce the fold-to-fold variation on the input values, thereby significantly reducing variability of the cross-validation error estimate of a classification rule.

SEQUENTIAL DESIGN AND ACTIVE LEARNING

In knowledge discovery, obtaining an informative training set is crucial to efficiently explore the data structure. Instead of a fixed sampling approach, an adaptive and sequential strategy can be more beneficial. This calls for *active learning*[75,84,85] in machine learning applications, whereby the learner actively selects data points from the predictor database $\mathcal{S}$ to be added to the training set. A standard approach in active learning using SVMs for classification is selection of the next data point with largest expectation of improvement.[86–88] Yu et al.[89] developed active learning using design ideas for regression models. A Bayesian sequential optimal design for sparse linear models[90] uses certain information-based loss functions.

Active learning is closely related to sequential experimental design. In sequential design, the data points are chosen sequentially, that is, $\boldsymbol{x}_{N+1}$ is selected based on $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ and their corresponding responses $y_1, y_2, \ldots, y_N$. There are two general approaches for generating sequential designs: stochastic approximation and variance optimization.

In *stochastic approximation* methods, explained here for $p = 1$ predictor, the xs are chosen such that x_N converges to the root of a certain function as $N \to \infty$. As a pattern classification method, stochastic approximation is widely used in many fields.[91] Consider the problem of stochastic root finding in sequential designs, that is, of estimating the value x at which $f(x) = E(y|x)$ attains a specified value α. Suppose the collected data are $(x_1, y_1), \ldots, (x_N, y_N)$. Robbins and Monro[92] proposed the stochastic approximation procedure

$$x_{N+1} = x_N - a_N(y_N - \alpha),$$

where $\{a_N\}$ is a prespecified sequence of positive constants. They also established the conditions under which x_N converges to the root of $f(x) - \alpha$. Joseph[93] developed an efficient Robbins–Monro procedure for binary data. Wu[94] proposed a stochastic approximation method for binary data known as the 'logit-MLE' method, in which $f(x)$ is approximated by a logit function $e^{(x-\mu)/\sigma}/(1 + e^{(x-\mu)/\sigma})$. Then, determination of x_{N+1} is by $x_{N+1} = \hat{\mu}_N + \hat{\sigma}_N \log\frac{\alpha}{1-\alpha}$, where $\hat{\mu}_N, \hat{\sigma}_N$ are maximum likelihood estimates of μ, σ based on $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$. Ying and Wu[95] show the almost sure convergence of x_N to the root irrespective of the function $f(x)$. Joseph et al.[96] improves Wu's logit-MLE method by giving more weight to data points closer to the root via a Bayesian scheme. When the design space $\mathcal{S}$ is multidimensional, multivariate stochastic approximation procedures have been studied by several researchers.[97–99] An overview of stochastic approximation methods and their theoretical properties can be found in Kushner and Yin[100] and Lai.[101]

In *variance optimization* methods for sequential design, a parametric model is postulated for the unknown function f. Points x are then chosen sequentially according to an optimality criterion such as A-, D-, or E-optimality. For example, Neyer[102] develops a sequential D-based design: the point x_{N+1} is chosen to maximize the determinant of the estimated information matrix. The root is solved from the final estimate of the function $f(x)$. Employing the D-criterion and a Bayesian analysis, Dror and Steinberg[103] offer a sequential design procedure for generalized linear models that can handle multiple predictors ($p > 1$) and can be applied for both fully sequential and group sequential settings. Lewi et al.[104] develop sequential optimal design for neurophysiology experiments by selecting data points to maximize the mutual information between the data and the unknown parameters of a generalized linear model. Sequential optimal design methods have been applied in many other areas including microarray[105,106] and psychophysical[107] studies.

Effectiveness of the variance optimization approach is model dependent: it performs best when the assumed model is the true model, and degrades with increasing deviation from the true model. Several methods on robust optimal design have been proposed to address the model uncertainties.[108,109] Understandably, the performance of the stochastic approximation method is not as good when the assumed model is correct. Deng et al.[110] introduce active learning via sequential design, which combines the advantages of both stochastic approximation and variance optimization methods.

The adaptive nature of sequential design and active learning is also embedded in the well-developed techniques of *response surface methodology* (RSM). Akin to techniques discussed earlier in this section, the objective of RSM is to identify settings of input variables $\boldsymbol{x}$ that optimize (typically, maximize) the expected response $f(\boldsymbol{x}) = E(y|\boldsymbol{x})$. As introduced by Box and Wilson,[111] RSM uses a sequence of designed experiments to attain an optimal response by fitting locally quadratic functions, with a new design generated at each step of a hill-climbing algorithm. For many industrial and biometric applications, RSM iterative learning can efficiently improve systems to maximize production. A comprehensive introduction is provided by Box and Draper.[112]

THE DESIGN HORIZON

As stated at the top of this article, experimental design seeks to optimize information garnered in scientific work. As indicated throughout this article the reach of design tools extends across the scientific spectrum. Meticulous data collection, that is, careful design, invariably improves the effectiveness of modeling and analysis techniques, and so enhances capacity to draw meaningful conclusions. With technological advance allowing ever larger amounts of data to be gathered on ever larger numbers of variables, design techniques are evolving to handle the increasingly complex data-collection and data-reduction problems entailed. To reiterate just two problems mentioned earlier, design matrix $\mathbf{X}$ choice for effective variable selection (with Lasso, SCAD, etc.), and SCV for building machine learners, are beginning to generate considerable interest, and offer many open avenues for innovation and improvement. The early efforts described here for extending design techniques to data mining and machine learning applications have the potential to engender a new interface between the two fields.

REFERENCES

1. Fisher RA. *The Design of Experiments*. Edinburgh, UK: Oliver and Boyd; 1937.
2. Elfving G. Optimum allocation in linear regression theory. *Ann Math Stat* 1952, 23:255–262.
3. Kiefer J. On the nonrandomized optimality and randomized nonoptimality of symmetrical designs. *Ann Math Stat* 1958, 29:675–699.
4. Kiefer J, Wolfowitz J. The equivalence of two extremum problems. *Can J Math* 1960, 12:363–366.
5. Box GEP, Draper NR. A basis for the selection of a response surface design. *J Am Stat Assoc* 1959, 54:622–654.
6. Atkinson AC, Donev AN, Tobias RD. *Optimum Experimental Designs, with SAS*. Vol. 34. Oxford Statistical Science Series. Oxford, UK:Oxford University Press; 2007.
7. Morgan JP, Wang X. Weighted optimality in designed experimentation. *J Am Stat Assoc* 2010, 105:1566–1580, (Supplementary materials available online).
8. Gilmour SG, Trinca LA. Optimal design of experiments for statistical inference. *Appl Stat* 2012, 61:1–25.
9. Dey A, Mukerjee R. *Fractional Factorial Plans*. Wiley Series in Probability and Statistics:

Probability and Statistics. New York: John Wiley & Sons; 1999.
10. Hedayat AS, Sloane NJA, Stufken J. *Orthogonal Arrays: Theory and Applications*. Springer Series in Statistics. New York: Springer-Verlag; 1999.
11. Cheng C-S. Orthogonal arrays with variable numbers of symbols. *Ann Stat* 1980, 8:447–453.
12. Mukerjee R. Universal optimality of fractional factorial plans derivable through orthogonal arrays. *Calcutta Stat Assoc Bull* 1982, 31:63–68.
13. Kuhfeld W. Orthogonal arrays. Available at: http://support.sas.com/techsup/technote/ts723.html. (Accessed January 2, 2012).
14. Sloane NJA. A library of orthogonal arrays. http://www2.research.att.com/~njas/oadir/. (Accessed January 2, 2012).
15. Xu H, Phoa FKH, Wong WK. Recent developments in nonregular fractional factorial designs. *Stat Surv* 2009, 3:18–46.
16. Fries A, Hunter WG. Minimum aberration 2^{k-p} designs. *Technometrics* 1980, 22:601–608.
17. Tang B, Deng L-Y. Minimum G_2-aberration for nonregular fractional factorial designs. *Ann Stat* 1999, 27:1914–1926.
18. Xu H, Wu CFJ. Generalized minimum aberration for asymmetrical fractional factorial designs. *Ann Stat* 2001, 29:1066–1077.
19. Xu H. Minimum moment aberration for nonregular designs and supersaturated designs. *Stat Sin* 2003, 13:691–708.
20. Jones B, Nachtsheim CJ. Efficient designs with minimal aliasing. *Technometrics* 2011, 53:62–71.
21. Sun D, Li W, Ye K. An algorithm for sequentially constructing non-isomorphic orthogonal designs and its applications. *Stat Appl* 2008, 6:144–158.
22. Angelopoulos P, Evangelaras H, Koukouvinos C, Lappas E. An effective step-down algorithm for the construction and the identification of nonisomorphic orthogonal arrays. *Metrika* 2007, 66:139–149.
23. Li Y, Deng L-Y, Tang B. Design catalog based on minimum *G*-aberration. *J Stat Plan Inference* 2004, 124:219–230.
24. Loeppky JL, Sitter RR, Tang B. Nonregular designs with desirable projection properties. *Technometrics* 2007, 49:454–467.
25. Xu H. Algorithmic construction of efficient fractional factorial designs with large run sizes. *Technometrics* 2009, 51:262–277.
26. Ryan KJ, Bulutoglu DA. Minimum aberration fractional factorial designs with large *N*. *Technometrics* 2010, 52:250–255, (Supplementary materials available online).
27. Tang B, Wu CFJ. A method for constructing supersaturated designs and its Es^2 optimality. *Can J Stat* 1997, 25:191–201.
28. Bulutoglu DA. Cyclicly constructed E(s^2)-optimal supersaturated designs. *J Stat Plan Inference* 2007, 137:2413–2428.
29. Nguyen N-K, Cheng C-S. New $E(s^2)$-optimal supersaturated designs constructed from incomplete block designs. *Technometrics* 2008, 50:26–31.
30. Jones BA, Li W, Nachtsheim CJ, Ye KQ. Model-robust supersaturated and partially supersaturated designs. *J Stat Plan Inference* 2009, 139:45–53.
31. Suen CY, Das A. $E(s^2)$-optimal supersaturated designs with odd number of runs. *J Stat Plan Inference* 2010, 140:1398–1409.
32. Tyssedal J, Samset O. Supersaturated designs of projectivity $P = 3$ or near $P = 3$. *J Stat Plan Inference* 2010, 140:1021–1029.
33. Marley CJ. *Screening Experiments Using Supersaturated Designs with Application to Industry* [Doctoral thesis]. Southampton, UK: University of Southampton; 2011, 136.
34. Georgiou SD. A general method for constructing supersaturated designs. Preprint.
35. Santner TJ, Williams BJ, Notz WI. *The Design and Analysis of Computer Experiments*. Springer Series in Statistics. New York: Springer-Verlag; 2003.
36. Fang K-T, Li R, Sudjianto A. *Design and Modeling for Computer Experiments*. Computer Science and Data Analysis Series. Boca Raton, FL: Chapman & Hall/CRC; 2006.
37. Loeppky JL, Sacks J, Welch WJ. Choosing the sample size of a computer experiment: a practical guide. *Technometrics* 2009, 51:366–376.
38. McKay MD, Beckman RJ, Conover WJ. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 1979, 21:239–245.
39. Morris MD, Mitchell TJ. Exploratory designs for computational experiments. *J Stat Plan Inference* 1995, 43:381–402.
40. Ye KQ , Li W, Sudjianto A. Algorithmic construction of optimal symmetric Latin hypercube designs. *J Stat Plan Inference* 2000, 90:149–159.
41. Husslage B, Rennen G, van Dam E, den Hertog D. Space-filling Latin hypercube designs for computer experiments. *Optim Eng* 2011, 12:611–630.
42. Grosso A, Jamali ARMJU, Locatelli M. Finding maximin latin hypercube designs by iterated local search heuristics. *Eur J Oper Res* 2009, 197:541–547.
43. Owen AB. Lattice sampling revisited: Monte Carlo variance of means over randomized orthogonal arrays. *Ann Stat* 1994, 22:930–945.

44. Tang B. Selecting Latin hypercubes using correlation criteria. *Stat Sin* 1998, 8:965–977.
45. Ye KQ. Orthogonal column Latin hypercubes and their application in computer experiments. *J Am Stat Assoc* 1998, 93:1430–1439.
46. Steinberg DM, Lin DKJ. A construction method for orthogonal Latin hypercube designs. *Biometrika* 2006, 93:279–288.
47. Sun F, Liu M-Q, Lin DKJ. Construction of orthogonal Latin hypercube designs with flexible run sizes. *J Stat Plan Inference* 2010, 140:3236–3242.
48. Bingham D, Sitter RR, Tang TB. Orthogonal and nearly orthogonal designs for computer experiments. *Biometrika* 2009, 96:51–65.
49. Lin CD, Mukerjee R, Tang B. Construction of orthogonal and nearly orthogonal Latin hypercubes. *Biometrika* 2009, 96:243–247.
50. Lin CD, Bingham D, Sitter RR, Tang B. A new and flexible method for constructing designs for computer experiments. *Ann Stat* 2010, 38:1460–1477.
51. Owen AB. Orthogonal arrays for computer experiments, integration and visualization. *Stat Sin* 1992, 2:439–452.
52. Tang B. Orthogonal array-based Latin hypercubes. *J Amer Stat Assoc* 1993, 88:1392–1397.
53. Tang B. A theorem for selecting OA-based Latin hypercubes using a distance criterion. *Commun Stat Theory Methods* 1994, 23:2047–2058.
54. Qian PZG. Sliced Latin hypercube designs. *J Am Stat Assoc*. In press.
55. Rennen G, Husslage B, Van Dam ER, Hertog DD. Nested maximin Latin hypercube designs. *Struct Multidiscip Optim* 2010, 41:371–395.
56. Loeppky JL, Moore LM, Williams BJ. Batch sequential designs for computer experiments. *J Stat Plan Inference* 2010, 140:1452–1464.
57. Qian PZG, Wu CFJ. Sliced space-filling designs. *Biometrika* 2009, 96:945–956.
58. Xu X, Haaland B, Qian PZG. Sudoku-based space-filling designs. *Biometrika* 2011, 98:711–720.
59. Hickernell FJ. A generalized discrepancy and quadrature error bound. *Math Comput* 1998, 67:299–322.
60. Fang K-T, Lin DKJ, Winker P, Zhang Y. Uniform design: theory and application. *Technometrics* 2000, 42:237–248.
61. Hill PDH. A review of experimental design procedures for regression model discrimination. *Technometrics* 1978, 20:15–21.
62. Dette H. Discrimination designs for polynomial regression on compact intervals. *Ann Stat* 1994, 22:890–903.
63. Dette H. Optimal designs for identifying the degree of a polynomial regression. *Ann Stat* 1995, 23:1248–1266.
64. Bingham DR, Chipman HA. Incorporating prior information in optimal design for model selection. *Technometrics* 2007, 49:155–163.
65. Beran R. Robust location estimates. *Ann Stat* 1977, 5:431–444.
66. Bhattacharyya A. On a measure of divergence between two statistical populations defined by their probability distribution. *Bull Calcutta Math Soc* 1943, 35:99–110.
67. Kakutani S. On equivalence of infinite product measures. *Ann Math* 1948, 49:214–224.
68. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc B* 1996, 58:267–288.
69. Breiman L. Better subset regression using the nonnegative garrote. *Technometrics* 1995, 37:373–384.
70. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Amer Stat Assoc* 2001, 96:1348–1360.
71. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Stat* 2004, 32:407–499.
72. Candes E, Tao T. The dantzig selector: statistical estimation when p is much larger than n. *Ann Stat* 2007, 35:2313–2351.
73. Deng X, Lin CD, Qian PZG. Designs for the Lasso. Tech. Rep. Madison, WI: Univeristy of Wisconsin-Madison; 2011.
74. Xing D, Wan H, Zhu Y. Optimal supersaturated design for variable selection via Lasso. Working paper, Purdue University, West Lafayette; 2011.
75. MacKay DJC. Information-based objective functions for active data selection. *Neural Comput* 1992, 4:590–604.
76. Cohn D. Neural network exploration using optimal experiment design. *Neural Networks*, 1996, 9:1071–1083.
77. Gilardi N, Faraj A. Design of experiments by committee of neural networks. In: *IEEE International Joint Conference on Neural Networks*. Budapest, Hungary: IEEE; 2004.
78. Seung H-S, Opper M, Sompolinsky H. Query by committee. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92. New York, NY: ACM; 1992, 287–294.
79. Haykin S. *Neural Networks: A Comprehensive Foundation*. Vol. 13. Prentice Hall; 1999.
80. Vapnik VN. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science. 2nd ed. New York: Springer-Verlag; 2000.
81. Staelin C. Parameter selection for support vector machines. 2002. Techn. Rep. HPL-2002-354. Haifa, Israel: HP Laboratories; 2002.
82. Stone M. Cross-validatory choice and assessment of statistical predictions. *J R Stat Soc B* 1974, 36:111–147.

83. Deng X, Qian PZG. Sliced cross-validation for efficient estimation of the error rate of a classification rule. Tech. Rep. Madison, WI: University of Wisconsin-Madison; 2011.
84. Cohn DA, Ghahramani Z, Jordan MI. Active learning with statistical models. *J Artif Intell Res* 1996, 4:129–145.
85. Fukumizu K. Statistical active learning in multilayer perceptrons. *Neural Netw IEEE Trans* 2000, 11:17–26.
86. Tong S, Koller D. Support vector machine active learning with applications to text classification. *J Mach Learn Res* 2002, 2:45–66.
87. Schohn G, Cohn D. Less is more: active learning with support vector machines. In: *Proceedings of the Seventeenth International Conference on Machine Learning*. San Francisco: Morgan Kaufmann; 2000, 839–846.
88. Campbell C, Cristianini N, Smola A. Query learning with large margin classifiers. In: *Proceedings of 17th International Conference on Machine Learning*. San Francisco: Morgan Kaufmann; 2000, 111–118.
89. Yu K, Jinbo Bi. Active learning via transductive experimental design. In: *Proceedings of the Twenty-Third International Conference on Machine Learning*. Pittsburgh: ACM; 2006, 1081–1088.
90. Seeger MW. Bayesian inference and optimal design for the sparse linear model. *J Mach Learn Res* 2008, 9:759–813.
91. Duda RO, Hart PE, Stork DG. *Pattern Classification*. 2nd ed. New York: Wiley Interscience; 2001.
92. Robbins H, Monro S. A stochastic approximation method. *Ann Math Stat* 1951, 22:400–407.
93. Joseph VR. Efficient Robbins–Monro procedure for binary data. *Biometrika* 2004, 91:461–470.
94. Wu CFJ. Efficient sequential designs with binary data. *J Am Stat Assoc* 1985, 80:974–984.
95. Ying Z, Wu CFJ. An asymptotic theory of sequential designs based on maximum likelihood recursions. *Stat Sin* 1997, 7:75–91.
96. Joseph VR, Tian Y, Wu CFJ. Adaptive designs for stochastic root-finding. *Stat Sin* 2007, 91:1549–1565.
97. Ruppert D. A Newton-Raphson version of the multivariate Robbins-Monro procedure. *Ann Stat* 1985, 13:236–245.
98. Wei CZ. Multivariate adaptive stochastic approximation. *Ann Stat* 1987, 15:1115–1130.
99. Spall JC, Member S. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans Autom Control* 1992, 37:332–341.
100. Kushner HJ, Yin GG. *Stochastic Approximation Algorithms and Applications*. New York: Springer-Verlag; 1997.
101. Lai TL. Stochastic approximation. *Ann Stat* 2003, 31:391–406.
102. Neyer BT. A d-optimality-based sensitivity test. *Technometrics* 1994, 36:61–70.
103. Dror HA, Steinberg DM. Sequential experimental designs for generalized linear models. *J Am Stat Assoc* 2008, 103:288–298.
104. Lewi J, Butera R, Paninski L. Sequential optimal design of neurophysiology experiments. *Neural Comput* 2009, 21:619–687.
105. Churchill GA. Fundamentals of experimental design for cDNA microarrays. *Nat Genet* 2002, 32(suppl Dec):490–495.
106. Durrieu G, Briollais L. Sequential design for microarray experiments. *J Am Stat Assoc* 2009, 104:650–660.
107. Rosenberger WF, Grill SE. A sequential design for psychophysical experiments: an application to estimating timing of sensory events. *Stat Med* 1997, 16:2245–2260.
108. Woods DC, Lewis SM, Eccleston JA, Russell KG. Designs for generalized linear models with several variables and model uncertainty. *Technometrics* 2006, 48:284–292.
109. Dror HA, Steinberg DM. Robust experimental design for multivariate generalized linear models. *Technometrics* 2006, 48:520–529.
110. Deng X, Joseph VR, Sudjianto A, Wu CFJ. Active learning through sequential design, with applications to detection of money laundering. *J Am Stat Assoc* 2009, 104:969–981.
111. Box GEP, Wilson KB. On the experimental attainment of optimum conditions. *J R Stat Soc B* 1951, 13:1–45.
112. Box GEP, Draper NR. *Response Surfaces, Mixtures, and Ridge Analyses*. 2nd ed. Wiley Series in Probability and Statistics. New York: John Wiley & Sons; 2007.

