Post-Fisherian Experimentation: from Physical to Virtual

C. F. Jeff Wu
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0205
(jeffwu@isye.gatech.edu)

Abstract

Fisher's pioneering work in design of experiments has inspired further work with broader applications, especially in industrial experimentation. Three topics in physical experiments are discussed: principles of effect hierarchy, sparsity, and heredity for factorial designs, a new method called CME for de-aliasing aliased effects, and robust parameter design. The recent emergence of virtual experiments on a computer is reviewed. Some major challenges in computer experiments, which must go beyond Fisherian principles, are outlined.

KEY WORDS: Aliased effects; Computer experiments; Conditional main effects; Effect heredity; Effect hierarchy; Effect sparsity; Kriging; Robust parameter design; Uncertainty quantification.

C. F. Jeff Wu is Professor and Coca Cola Chair in Engineering Statistics, School of Industrial and Systems Engineering, Georgia Institute of Technology, 765 Ferst Drive, NW, Atlanta, GA 30332-0205 (E-mail: jeffwu@isye.gatech.edu). Thanks go to the Fisher lectureship committee for the invitation to deliver the Fisher lecture at the 2011 Joint Statistical Meetings and my mentors Peter Bickel and George Box. The author is grateful to T. Dasgupta, M. Hamada, V. R. Joseph, P. Z. Qian and two referees for their helpful comments on an earlier draft of the paper. This work is supported by NSF grants DMS 1007574 and DMS 1308424 and DOE grant DE-SC0010548.

1 Introduction

In October 1919 R.A. Fisher got his first regular job at Rothamsted Experimental Station. According to the recollection of the station director Sir John Russell, he was looking for a young mathematician "who would be prepared to examine our data and elicit further information that we had missed" (Box, 1978, p. 96). By 1926, in a mere seven years, Fisher had invented the ANalysis Of VAriance (ANOVA) and Design Of Experiments (DOE). The rest is history, as the two methods are viewed as among the most important scientific developments in the first half of the 20th century.

Fisher's work in DOE was inspired by the agricultural experiments at Rothamsted. It had made inroads to industrial experiments, particularly in light industry such as textiles. Its wide-spread applications in industry came after World War II because of the needs of massive industrialization. The main difference between agricultural and industrial experiments is that the former takes a longer time, needs more planning and is subject to uncontrollable variations in the field while the latter can be conducted in the lab or factory floor and the duration is shorter. There are also more input factors in industrial experiments as their purpose is often the improvement or optimization of processes. Unlike agricultural experiments which use blocking more extensively, industrial experiments have more control variables and factorial designs are used to conduct such experiments. Even though the basic ideas of factorial designs were developed before the war by Fisher, Yates and Finney, their big push came after the war, especially the work of the Wisconsin School led by George Box. Then the complexity of relationships between factors became a new research issue. When Mike Hamada and I were preparing our book (Wu and Hamada, 2000), we were looking for principles that govern the relationships between factorial effects of various orders analogous to the principles of replication, blocking and randomization in Fisher's

development of DOE. In Section 2, I will describe them as the principles of *effect hierarchy, sparsity and heredity.* In tracing their historical origins or precedents for preparing the Fisher Lecture, I found with satisfaction and somewhat to my surprise the early work of Yates (1935, 1937), who was closely associated with Fisher and his DOE work. Some impacts of these principles will also be discussed.

Effect aliasing is a basic concept and necessary evil in fractional factorial design. Since the pioneering work of Finney (1945), it has been universally accepted that, if two factorial effects are (fully) aliased, they cannot be disentangled unless more data are collected. In 1988 I asked the following question: can two aliased effects be "de-aliased" without adding more runs? On the face of it, this did not seem possible. By looking into the notion of interactions more closely, it will be shown in Section 3 that this is indeed possible. The main underlying idea is to view a two-factor interaction as the difference between two Conditional Main Effects. This new method is thus termed *CME analysis.* A successful application of this method will be illustrated with an industrial experiment at General Motors of Canada.

Another major development after World War II is robust parameter design. It was pioneered by G. Taguchi (1987) largely based on engineering concepts and his many years of experience working with industries in Japan. It has made an impact in the practice of quality and productivity improvement. His new paradigm has also led to the rejuvenation of research on design and analysis of experiments, especially in the period 1985-2000. Section 4 gives a brief description of the methodology and discusses their major deviations and differences from the traditional strategy in conducting physical experiments.

In the long history of DOE, physical experiments have played a dominant role for its first 70-80 years. As the needs over time changed, physical experiments took

different forms. It started with agriculture in its founding days, moved to process industries as witnessed by Box's work in 1950's at Imperial Chemical Industries (Box, 2013), and to manufacturing as witnessed by Taguchi's (1986, 1987) work on robust parameter design for variation reduction and quality improvement. The material in Sections 2-4 gives a personalized glimpse of some work on physical experiments in this long period. In the last decade, the interest and focus have gradually shifted from physical to virtual experiments. Because of the advances in complex mathematical modeling and information technology, virtual experiments on a computer, or called computer experiments, have become popular in engineering and scientific investigations. Computer simulations can be much faster or less costly than running physical experiments. Furthermore, physical experiments can be difficult to conduct as in the detonation of explosive materials or infeasible when rare events like landslides or hurricanes are observed. Section 5 gives a generic description of computer experiments and outlines some challenges they face. There are two major approaches to the modeling and computation of computer experiments: stochastic and numerical. Some of the difficulties for each approach are discussed. Connections to uncertainty quantification (UQ), which originated in applied mathematics, are also made. Concluding remarks are given in Section 6.

2 Guiding principles for factorial designs

The following three principles, which govern the relationships among factorial effects, are given in Wu and Hamada (2000, 2009):

(i) *Effect hierarchy*: Lower order effects are more likely to be important than higher order effects; effects of the same order are equally likely to be important.

(ii) *Effect sparsity*: The number of relatively important effects is small.

(iii) *Effect heredity*: For an interaction to be significant, at least one of its parent main effects should be significant.

More discussion and references on the three principles can be found in Sections 4.6 and 9.5 of Wu and Hamada (2009). Here we focus on their historical connections and impacts. The term "effect hierarchy" was coined in Wu and Hamada (2000, 2009) but the concept was implicit in design textbooks for decades. This concept was in fact mentioned explicitly in the pioneering work of Yates on factorial designs. In Yates (1935, p. 209), he said "From physical considerations and practical experience, (interactions) may be expected to be small in relation to error ..." Yates (1937, p. 18) stated this concept more explicitly "... higher-order interactions ... are usually of less interest than the main effects and interactions between two factors only." Yates used this concept to guide data analysis. In addition to helping data analysis, the more precise version given here can be used to justify the choice of "optimal" fractions of factorial designs such as the maximum resolution criterion and the minimum aberration criterion (Wu and Hamada, 2009, Section 5.2). It has also been used in justifying other optimal design criteria (see Mukerjee and Wu, 2006).

The term "effect sparsity" was coined by Box and Meyer (1986) but the concept went further back to Box and Hunter (1961). On page 314 of the 1961 paper, they stated "In some situations the total number of variables k is large, but only a few (of them) are expected to have any effect." A similar idea was stated in the last paragraph of page 341 of the same paper. The concept can be used to justify the use of screening designs such as 2^{k-p} designs with large fractions. It can also be used to guide data analysis. This concept has its precedent in quality engineering. The Pareto chart (or Pareto histogram) is one of the seven tools in quality control, sometimes

referred to as the magnificent seven (Ishikawa, 1971). The quality guru Juran found out from his consulting experience that, in many quality investigations, a *vital few* of the defects account for most of the total effect while the remaining ones (the *trivial many*) account for little of the total effect. The Pareto histogram is obtained by rearranging the histogram of defects in the descending order of frequencies. He argued that quality investigation should focus on the top few defects or causes. While this is technically very easy, its underlying philosophy is deep. Juran developed this concept in the late 1940's and gave it the name "Pareto principle". For a fascinating account, read Juran (1974, Section 2, pages 16-18). Given this historical background, we may call effect sparsity the *Pareto principle in DOE*.

Effect heredity was proposed by Hamada and Wu (1992) in the context of model selection for factorial experiments whose effects have complex aliasing relationships. Its original purpose was to rule out incompatible models in model search. The version stated here is called *weak heredity* by Chipman (1996) because it allows either one of the parent factors to be present in the model. If both parent factors should be present, it is called *strong heredity*. Strong heredity is closely related to the notion of "marginality" (Nelder, 1977; McCullagh and Nelder, 1989). But the purpose of marginality was different, i.e., to maintain invariance of the response surface under scaling and translation of the factors of an experiment. When Mike and I wrote our 1992 paper, we had no idea if this concept was already available in the design literature. To my pleasant surprise, it was stated in Yates (1937, p. 12) "... factors which produce small main effects usually show no significant interactions." As in the case of effect hierarchy, Yates used it to guide data analysis. The more precise version stated here and those in Chipman (1996) can be used for model selection and data analysis. Since effect heredity is used to rule out incompatible models, it appears natural that it should have wider use beyond designed experiments. Lately

it has been successfully applied to variable selection in general regression settings. Its generic use can be described as follows. For a given model search algorithm, effect heredity is imposed in the search procedure. The first two such applications are to LARS (Efron et al., 2004) by Yuan, Joseph, and Lin (2007) and to the nonnegative garrote (Breiman, 1995) by Yuan, Joseph and Zou (2009).

Each of the three principles can be given a Bayesian formulation. A main advantage of the Bayesian approach is that it can define the principles in a quantitative way. Consider, for example, the effect sparsity principle. If the prior probability of a factor being significant is small, it indicates a stronger degree of sparsity. Therefore the prior probability gives a quantitative measure of sparsity. Similarly, the weak and strong effect heredity principles have a Bayesian formulation and the degree of inheritance can be determined by choosing appropriate values in the prior parameters. For details, see Chipman (1996) or Section 9.5 of Wu and Hamada (2009).

Let me conclude this section by giving an example of analysis that benefits from use of these principles. It is well known that many of the nonregular factorial designs have complex aliasing among their factorial effects. The 12-run Plackett-Burman design with 11 factors is probably the most "notorious" because each of its two-factor interactions is partially aliased (with coefficient 1/3 or -1/3) with any main effects not present in the two-factor interaction (Wu and Hamada, 2009, Section 9.1). The aliasing relationship is called complex because each main effect has 45 ($= \binom{10}{2}$) partial aliases. This complex relationship was viewed as a disadvantage of these designs in the standard literature (up to 1992). Therefore they were mostly used for the purpose of factor screening, i.e., to estimate the main effects only. Hamada and Wu (1992) argued that, in view of effect sparsity, the number of terms in a selected model in actual data analysis is much smaller than what the combinatorial computation on the

size of model search would indicate (prior to analysis). Thus effect sparsity can be used to reduce the complexity of effect aliasing. Then they employed effect heredity to drastically reduce the space for model search by eliminating models that violate the effect heredity relationship. The strategy was successfully illustrated on two real experiments, one using a 12-run Plackett-Burman design with seven factors and the other using a 36-run orthogonal array with seven three-level factors and one two-level factor. Details can be found in Chapter 9 of Wu and Hamada (2009).

3 Conditional main effect (CME) analysis

Let me start by briefly reviewing the concept of effect aliasing. Consider the 2^{4-1} design given in Table 1 with eight runs and four factors A, B, C, D. It is a half fraction of the 2^4 design with $\mathbf{I} = ABCD$ as its defining relation. The two-factor interactions (henceforth abbreviated as 2fi's) AB and CD are given in the last two columns of the table. One can see that the two columns share the same column vector (known as a contrast in DOE). This is a consequence of the relationship $\mathbf{I} = ABCD$. The 2fi's AB and CD are said to be *aliased* (Finney, 1945) because they represent the *same* contrast. In group-theoretic terms, both of them belong to the same coset. As in Wu and Hamada, we will call them *fully* aliased if there is a need to distinguish it from the notion of partial aliasing, which will be used below. See also Wu and Hamada (2009, pages 290 and 422). For the 2^{4-1} design with $\mathbf{I} = ABCD$, there are seven degrees of freedom, each of which is associated with a coset. Since each coset represents one degree of freedom, it is not possible to distinguish AB and CD.

It has been widely accepted since the pioneering work of Finney (1945) that aliased effects cannot be disentangled. In 1988 I raised the question: is it possible for aliased effects in two-level designs to be "de-aliased" without adding runs? The key to unlock

this puzzle is to revisit the definition of interaction. Suppose there are two factors A and B, each at two levels denoted by $+$ and $-$. The standard definition for the $A \times B$ interaction in design textbooks is

$$\begin{aligned} \mathrm{INT}(A,B) &= \frac{1}{2}\{\bar{y}(A+|B+)+\bar{y}(A-|B-)\} \\ &\quad -\frac{1}{2}\{\bar{y}(A+|B-)+\bar{y}(A-|B+)\}, \end{aligned} \tag{3.1}$$

where $\bar{y}(A+|B+)$, $\bar{y}(A-|B+)$, $\bar{y}(A+|B-)$ and $\bar{y}(A-|B-)$ are respectively the average value of the response y at the four settings $A+B+$, $A-B+$, $A+B-$ and $A-B-$. Following Wu and Hamada (2009, Section 4.3.2), we consider an alternative and equivalent definition using the notion of conditional main effects. Specifically, we define the *conditional main effect* (abbreviated as cme) of B at the $+$ level of A as

$$\mathrm{CME}(B|A+) = \bar{y}(B+|A+)-\bar{y}(B-|A+). \tag{3.2}$$

Thus we can interpret $\mathrm{CME}(B|A+)$ as the main effect of B conditional on A being at the $+$ level. The definitions of $\mathrm{CME}(B|A-)$, $\mathrm{CME}(A|B+)$, and $\mathrm{CME}(A|B-)$ can be similarly made. Then it is straightforward to show that

$$\begin{aligned} \mathrm{INT}(A,B) &= \frac{1}{2}\{\mathrm{CME}(B|A+)-\mathrm{CME}(B|A-)\} \\ &= \frac{1}{2}\{\mathrm{CME}(A|B+)-\mathrm{CME}(A|B-)\}. \end{aligned} \tag{3.3}$$

These relationships suggest that we can view the two cme's $\mathrm{CME}(B|A+)$ and $\mathrm{CME}(B|A-)$ as two components of the interaction $\mathrm{INT}(A,B)$. More generally we should consider an interaction together with its two parent main effects in a three-dimensional (3d) space. In the traditional approach this space consists of the three orthogonal components: $\mathrm{INT}(A,B)$ and its two main effects A and B. Using the cme concept, this space can be reparameterized as: $\mathrm{CME}(B|A+)$, $\mathrm{CME}(B|A-)$ and the main effect A of their conditioning factor. These three effects are also mutually

Table 1: A 2^{4-1} design with $\mathbf{I} = ABCD$

A	B	C	D	AB	$= CD$
$-$	$-$	$-$	$-$	$+$	$+$
$-$	$-$	$+$	$+$	$+$	$+$
$-$	$+$	$-$	$+$	$-$	$-$
$-$	$+$	$+$	$-$	$-$	$-$
$+$	$-$	$-$	$+$	$-$	$-$
$+$	$-$	$+$	$-$	$-$	$-$
$+$	$+$	$-$	$-$	$+$	$+$
$+$	$+$	$+$	$+$	$+$	$+$

orthogonal but, unlike the first case, their corresponding vectors do not have the same length. We can also reverse the roles of A and B to get a third parametrization of the 3d space. Returning to Table 1, we can see that its six columns form a five-dimensional space because of the relationship $AB = CD$. Consider now Table 2. Its columns A, $B|A+$ and $B|A-$ form a 3d space, where $B|A+$ and $B|A-$ are the shorthand notation for CME($B|A+$) and CME($B|A-$). Similarly C, $D|C+$ and $D|C-$ in Table 2 form another 3d space. By amalgamating these two spaces, their joint space has five dimensions because it is the same as the one in Table 1. However the six components of this 5d-space are not mutually orthogonal, i.e., some of them are partially aliased. Because no effects in this space are fully aliased, we can use forward-type variable selection to identify significant effects among the six candidate effects. Therefore we are able to "de-alias" the pair AB and CD through some cme's in the 5d space. We call this approach the *CME analysis*. In this problem *non-orthogonality is the saving grace*. By contrast, the traditional approach cannot resolve aliasing because it is based on Table 1, which has five orthogonal components but also a pair of fully aliased components.

So far the CME analysis is described in the context of the parametrization given

Table 2: A 2^{4-1} design with its eight effects

A	B	C	D	$B\|A+$	$B\|A-$	$D\|C+$	$D\|C-$
$-$	$-$	$-$	$-$	0	$-$	0	$-$
$-$	$-$	$+$	$+$	0	$-$	$+$	0
$-$	$+$	$-$	$+$	0	$+$	0	$+$
$-$	$+$	$+$	$-$	0	$+$	$-$	0
$+$	$-$	$-$	$+$	$-$	0	0	$+$
$+$	$-$	$+$	$-$	$-$	0	$-$	0
$+$	$+$	$-$	$-$	$+$	0	0	$-$
$+$	$+$	$+$	$+$	$+$	0	$+$	0

in Table 2. It can be extended to general two-level fractional factorial designs. The generic idea is to identify all aliased pairs of 2fi's and their corresponding cme's. Then use variable selection to identify significant effects from a set of candidate effects consisting of some main effects, clear 2fi's, and cme's that correspond to the aliased 2fi's. (Note that a *clear* 2fi is not aliased with the main effects and any other 2fi's; see p. 214 of Wu and Hamada (2009).) The trick lies in the choice of a candidate set of effects and also in the variable selection strategy. Since this involves further algebraic development, a systematic analysis strategy based on the CME concept will be given in Su and Wu (2014). One example of the choice of candidate set is given in the following analysis.

The CME analysis is now illustrated with some data from GM of Canada (Brajac and Morey, 1987). This was the first data set I applied the method to in 1988. A simulation was run to mimic an assembly subprocess called marriage. A vehicle component such as an engine or axle was carried by an automatic guided vehicle (AGV) to the marriage station to be fastened to a vehicle underbody. There were six factors that could affect the outcome of this operation as measured by its throughput within a 40-hour period. The simulation study was used to understand which of the

Table 3: Factors and levels, car marriage simulation experiment

Factor	Level	
	$-$	$+$
A. No. of lanes in brake cell	3	4
B. % of cars with ABS	0	100
C. lane selection logic	FIFO	free flow
D. No. of Automatic Guided Vehicles	24	34
E. % repair in marriage	8	16
F. marriage base cycle time	124	$124 + 29$

factors and their levels (see Table 3) had significant effects on the throughput. A 16-run 2^{6-2} design with the defining relations $\mathbf{I} = ABCE = ABDF = CDEF$ was chosen for the experiment. It is easy to show that this is a resolution IV design and that each of its 2fi's is aliased with one or two other 2fi's. This is an ideal case to show the potential advantage of the CME analysis because all of its 2fi's are aliased and thus the traditional approach cannot be used to estimate interactions with no ambiguity.

The design matrix and data are given in Table 4. The run order of the experiment was randomized. Each response value in the table is called job loss because it is obtained by subtracting the throughput value (from the simulation run) from a target value of 2880.

First we tried the traditional approach by using the half-normal plot to identify significant effects. From Figure 1, it is clearly seen that the main effect E (repair rate) is the most significant and much larger than the other effects. This is obvious because repair rate is expected to heavily influence the throughput (or equivalently the job loss). It is followed by two other main effects C and A. The R^2 value for the three effects is 97.6%. If the next effect CF is added to E, C, A, the resulting model, called Model 1, has a slightly larger R^2 (= 98.29%). The p values for C, A and CF

Table 4: Design matrix and data, car marriage simulation experiment

A	B	C	D	E	F	y
$-$	$-$	$-$	$-$	$-$	$-$	13
$+$	$+$	$-$	$-$	$-$	$-$	5
$-$	$-$	$+$	$-$	$+$	$-$	69
$-$	$-$	$-$	$+$	$-$	$+$	16
$+$	$-$	$+$	$+$	$-$	$-$	5
$+$	$-$	$+$	$-$	$-$	$+$	7
$+$	$-$	$-$	$+$	$+$	$-$	69
$+$	$-$	$-$	$-$	$+$	$+$	69
$-$	$+$	$+$	$+$	$-$	$-$	9
$-$	$+$	$+$	$-$	$-$	$+$	11
$-$	$+$	$-$	$+$	$+$	$-$	69
$-$	$+$	$-$	$-$	$+$	$+$	89
$+$	$+$	$+$	$-$	$+$	$-$	67
$+$	$+$	$-$	$+$	$-$	$+$	13
$-$	$-$	$+$	$+$	$+$	$+$	66
$+$	$+$	$+$	$+$	$+$	$+$	56

are respectively 1.9%, 2.2% and 5.6%. This model has four terms. Furthermore, even if CF is significant, we cannot give it a good interpretation because CF is aliased with DE (which follows from the relation $\mathbf{I} = CDEF$). The CME analysis can be handily used here. First we need to find a candidate set of effects for performing variable selection. This set consists of all the six main effects and all the eight cme's associated with DE and CF. Stepwise or forward variable selection identifies three effects: E, $C|F+$, and A, which form Model 2. It has three terms with $R^2 = 98.26\%$. The p values for $C|F+$ and A are respectively 0.37% and 1.8%, which are more significant than C, A and CF in Model 1. Clearly Model 2 is better than Model 1.

Finally we use this result to demonstrate the practical meaning of a significant conditional main effect. The $C \times F$ interaction plot is given in Figure 2. It is seen that the conditional main effect of C at $F+$ (the solid line) is much larger than that

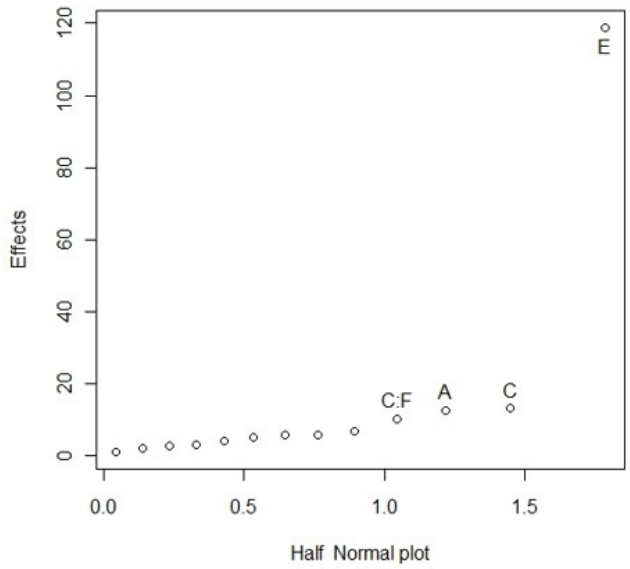

Figure 1: Half-normal plot of effects, car marriage simulation experiment

at $F-$. So what is the engineering meaning of a significant $C|F+$? It says that the lane selection logic (factor C) has a significant effect on job loss for larger cycle time ($F+$) but not for smaller cycle time ($F-$). It was explained in the original GM study that the main difference between $F-$ and $F+$ is that the latter added another 29 seconds of cycle time by completing the remaining work in one of the buffer positions outside the marriage station. It is only in this situation that the choice of lane logic (first-in-first-out or free flow) makes a difference on job loss. By comparison, it would be difficult to give a meaningful engineering interpretation of a significant CF interaction in view of its definition in (3.1), even if CF were a clear effect. Another point to note is that the CME analysis can be applied even when the 2fi's in a design are clear. This is particularly useful when the conditional main effects are more meaningful or interpretable than the two-factor interactions (defined

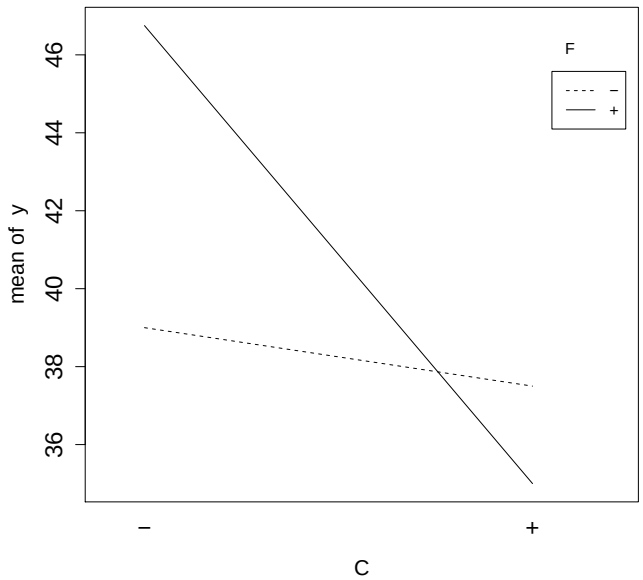

Figure 2: $C \times F$ interaction plot, car marriage simulation experiment

in (3.1)) for a given problem.

The concept of conditional main effects has been used in other statistical modeling. For example, Wu and Hamada (2009, Section 7.9.1) used it in nested effects modeling for experiments with sliding levels. It also has applications beyond engineering. Two examples are given here for illustration. If varieties differ in their response to environmental change, there is a genotype $\times$ environment interaction. Denote the genotype by A and the environment by B, each assumed at two levels. Then their interaction can be defined as in (3.1). If the breeder cares only about the difference between the two varieties across the environment, then we can use the main effect of A to measure this difference. If, however, the difference of responses in the two varieties varies substantially with the environment, then the conditional main effect of A given $B+$ (or $B-$) may be more relevant because of the need to develop breeds that adapt locally

to the specific environment. More discussion on genotype $\times$ environment interactions can be found in Lynch and Walsh (1998). See, especially, Figures 6.6, 6.7 and 22.1. The second example is taken from social sciences. Consider a study on a population of schools. There are two experimental factors (interventions): a performance-based bonus scheme for teachers (A), and a quarterly review by a team of external experts (B). Each factor has two levels: 0 (no intervention) and 1 (intervention applied). In this case, the conditional main effect of B given A at level 0 is of more interest than the unconditional main effect of B. The schools that do not receive the bonus scheme are likely to be dissatisfied or lack motivation in comparison to the schools who receive the scheme. Consequently, the former group is likely to benefit more from a quality review as compared to the latter. Thus, this conditional main effect represents how much the quality review can excite a group of less motivated teachers.

Finally, an interesting question (rasied by a referee) is whether or how the adoption of the CME analysis would have any implications on the design of the study.

4 Robust parameter design

Robust parameter design is another major development after World War II. It is less well known in the statistical world than other major developments like response surface methodology. This is primarily due to its use of engineering jargon and the sometimes confusing use of statistical concepts by its protagonist G. Taguchi. Taguchi classified the input factors of a system (product or process) into two types: control and noise. Control factors are those whose values remain fixed once they are chosen in an experiment. They can include design parameters in product or process design such as part dimensions and heat treatment time. Noise factors are those that are hard to control during the normal process or use condition. They can be variations around

nominal values in a mechanical design or temperature fluctuation in an oven. The gist of robust parameter design (or simply robust design) is to reduce the response variation of a system by choosing the settings of some control factors to make it less sensitive (i.e., more robust) to noise variation. The methodology has been widely used in quality and productivity improvement and has become part of the toolkit in quality engineering. The idea is illustrated in Figure 3, where Y represents the response, X the control factors and Z the noise factors. In traditional design (see Figure 3(a)), the response variation is reduced by reducing the noise variation of the input Z while the nominal value of X is fixed at X_1. Recognizing that the reduction of noise variation can be time-consuming or costly, Taguchi suggested an innovative way to reduce response variation, which is depicted in Figure 3(b): keep the noise variation intact but move the nominal value of X from X_1 to X_2. In order for this to be effective, the strategy must exploit interactions between the control and noise factors. Therefore the control $\times$ noise interactions, denoted by CN, play a pivotal role in robust design. Details on robust design can be found in books by Taguchi (1986, 1987), Montgomery (2005) and Wu and Hamada (2009).

Because of its engineering origin and the goal of variation reduction, some key aspects of robust design are quite different from the traditional strategy. Three will be discussed here. First, the emphasis shifts from location effects as in Fisherian ANOVA to dispersion effects because of its focus on variation reduction. This can be seen in Taguchi's ubiquitous two-step procedures for parameter design optimization. (Several versions of the two-step procedure can be found in Taguchi (1987) and Wu and Hamada (2009).) In most cases the first step is to reduce a dispersion measure like maximizing the signal-to-noise (SN) ratios he proposed for various classes of problems. Only in the second step of adjustment is the location effect (e.g., sample mean) considered. This change of focus has implications in the analysis and opti-

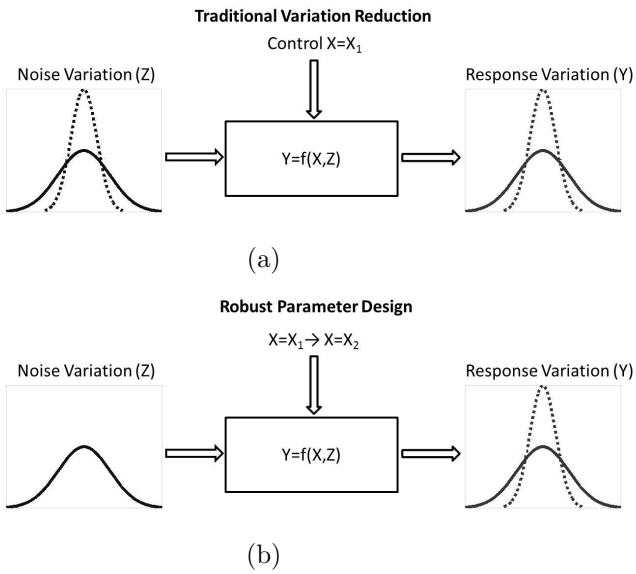

Figure 3: Variation reduction through robust parameter design

mization strategies, which will be briefly discussed in the third point below. Second, the major role played by the CN interactions make them as important as the control main effects C and the noise main effects N. This violates the intent of the effect hierarchy principle. It has a major influence on the development of optimal design theory for robust design experiments. Thanks to the effect hierarchy principle which treats effects of the same length equally, design of traditional factorial experiments can make use of Galois theory through the defining contrast subgroup that defines the fractional factorial design. This is no longer true in designing fractional factorial experiments for robust design applications. See, for example, Wu and Zhu (2003). Finally, Taguchi's use of performance measures (e.g., his SN ratios) for system optimization often leads to a different modeling strategy. The prevailing approach in statistics is to find a good model to fit the response data and then optimize the performance measure based on the identified model. Taguchi's preferred approach is to model *directly* the performance measure like SN ratios and then perform optimization

based on the fitted model. In Wu and Hamada (2009, Chapter 11), these two approaches are called respectively the response modeling and the performance measure modeling. If a performance measure modeling leads to spurious interactions, then it may not work well. Otherwise, the direct modeling of a measure to be optimized can be appealing to users, even when the response modeling enjoys some advantages in statistical efficiency. An in-depth comparison of the two approaches will have benefits beyond robust design because the same issue also arises in other contexts.

5 Virtual experiments on a computer: modeling and computation

Because of the rapid advance in high-fidelity mathematical modeling and fine-scale computation, it has become practical to use computer simulations to mimic real-world phenomena. I will first focus on deterministic simulations, which are called deterministic computer experiments in the statistical literature. A simulation is called deterministic if the same input values will lead to the same output. Consider, for example, the use of computer simulations to design a heat exchanger (Qian et al., 2006). Input factors that can affect the performance of a heat exchanger include cell topology, dimensions, wall thickness, conductivity of solid, temperature of heat source, and so forth. It is common to use a computational fluid dynamics (CFD) solver such as FLUENT to solve the heat transfer problem. In this case the simulation output of interest (or the response) is the maximum total heat transfer. For the purpose of heat exchanger design, a higher response value is desired. The engineering problem is to use these simulations over a variety of the input values to identify combinations of the input factors that maximize the response. Denote the input values by a vector $\mathbf{x} = (x_1, \ldots, x_p)$ and the output by y. The relationship between y and $\mathbf{x}$ is given

$y = g(\mathbf{x})$, where g is often a complicated function. The computation of $g(\mathbf{x})$ can be done in a variety of ways, depending on the nature of the problem. One can use finite element analysis (FEA) to solve a set of partial differential equations (pde's) such as in FLUENT. Alternatively, one can use discrete-event simulation (DES) to analyze dynamic, stochastic systems which involve queuing systems, e.g., queue of jobs to be processed on a printer. A good example is the car marriage simulation experiment in Section 3.

The computation of $g(\mathbf{x})$ can be precise but also time-consuming. If a large number of simulations over the $\mathbf{x}$ combinations need to be performed, which is often required in product design and optimization, it will be infeasible to use computer simulations to complete the task. Instead a *surrogate model* (or *meta model*) can be built by taking the results of the simulations as input values. There are broadly two approaches to building surrogate models: stochastic and numerical. Typically a surrogate model has an explicit or computationally efficient form for the relationship between y and $\mathbf{x}$. Thus it can be used to evaluate the y value for any $\mathbf{x}$ outside the chosen sites $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ much more expediently than the original simulations. This surrogate model is also referred to as an *emulator* as opposed to the original *simulator*. The availability of an emulator gives the investigator the capability to study the function g over a wide portion of the input region. If the predicted values from the surrogate model do not conform to the intuition or expectation of the investigator, he/she can return to the simulator to perform more simulations at the suspicious sites. If these results are quite different from the predicted values by the emulator, the emulator can be updated based on the expanded set of values from the simulation. Thanks to the interpolating property of an emulator like kriging (see below), the observed discrepancy at the $\mathbf{x}$ sites in the second round of the simulation will be reduced to zero. The relationship between the simulator and the emulator is depicted

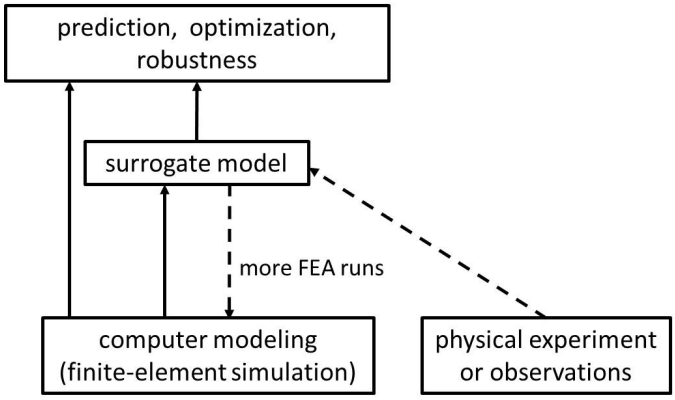

Figure 4: Statistical surrogate modeling of computer experiments

in the left portion of Figure 4. The two-way flow enables simulation and emulation to be conducted iteratively. A physical experiment or field observations can also be brought into the picture (lower-right portion of Figure 4). The physical observations together with the simulation results can be used jointly in training a surrogate model. A good example is the problem of estimating the calibration parameters in a computer model by using physical observations (Kennedy and O'Hagan, 2001). The objective of the investigation is depicted in the top box of Figure 4. It may be prediction, optimization or robustness (Santner et al., 2003). It can be based on the original simulation results or on the surrogate model results, especially if the former does not have enough data.

Kriging is the most popular stochastic approach to modeling computer experiments. The relationship between y and $\mathbf{x}$ can be described by the following model:

$$y(\mathbf{x}) = \mu(\mathbf{x}) + Z(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^p, \tag{5.1}$$

where $\mu(\mathbf{x}) = \sum_{i=0}^{m-1} \beta_i f_i(\mathbf{x})$, $f_i(\mathbf{x})$ are some known functions, β_i's are some unknown parameters, $Z(\mathbf{x})$ is a stationary stochastic process with mean 0 and covariance function $\mathrm{Cov}\{Z(\mathbf{x}+\mathbf{h}), Z(\mathbf{x})\} = \sigma^2 R(\mathbf{h}; \boldsymbol{\theta})$, and $\boldsymbol{\theta}$ is a vector of unknown correlation

parameters. The linear model part μ in (5.1) captures the global trend, whereas the stochastic model part Z in (5.1) captures the local trend. The stochastic model in (5.1) can also be given a Bayesian interpretation, i.e., it imposes a Bayesian prior on the space of deterministic functions $g(x)$. See Section 2.3 of Santner et al. (2003) for details. Suppose we perform computer experiment at n sites $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and obtain the corresponding function values $\mathbf{y} = (y_1, \ldots, y_n)'$. The predictor of y at any $\mathbf{x}$ can be obtained as follows:

$$\hat{y}(\mathbf{x}) = \mathbf{f}(\mathbf{x})'\hat{\boldsymbol{\beta}} + \mathbf{r}(\mathbf{x})\mathbf{R}^{-1}(\mathbf{y} - \mathbf{F}\hat{\boldsymbol{\beta}}), \tag{5.2}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{F}'\mathbf{R}^{-1}\mathbf{F})^{-1}\mathbf{F}'\mathbf{R}^{-1}\mathbf{y}$. Here $\mathbf{f}(\mathbf{x}) = (f_0(\mathbf{x}), \ldots, f_{m-1}(\mathbf{x}))'$ is the vector of known functions at $\mathbf{x}$, $\mathbf{F}$ is the $n \times m$ regression model matrix defined as $\mathbf{F} = (\mathbf{f}(\mathbf{x}_1), \ldots, \mathbf{f}(\mathbf{x}_n))'$, $\mathbf{R}$ is the $n \times n$ correlation matrix whose ijth element is $R(\mathbf{x}_i - \mathbf{x}_j; \boldsymbol{\theta})$, and $\mathbf{r}(\mathbf{x}) = (r_1(\mathbf{x}), \ldots, r_n(\mathbf{x}))'$ is the $n \times 1$ correlation vector with $r_i(\mathbf{x}) = R(\mathbf{x} - \mathbf{x}_i; \boldsymbol{\theta})$. The correlation parameters $\boldsymbol{\theta}$ are usually estimated from data using the maximum likelihood method. An important property for the kriging predictor $\hat{y}(\mathbf{x})$ in (5.2) is that it is an *interpolator*, i.e., $\hat{y}(\mathbf{x}_i) = y_i$. It is especially desirable for deterministic computer simulations.

The kriging technique originated in the spatial statistics literature but the mathematical work goes back earlier. See Cressie (1990) for its history. Its use in computer experiments started in the 1980's, especially due to the work of Sacks, Ylvisaker, Mitchell, Welch, etc. See the book by Santner et al. (2003) for references. However, its applications to computer experiments have some major differences from those in spatial statistics. Because a simulation problem can have a large number of input factors, the input dimension for computer experiments can be much higher than two or three (which is common in spatial statistics). Therefore variable selection becomes an important issue. Also the nature of the input factors can be quite different, e.g.,

temperature, dimension, and velocity. Most of the spatial statistics studies are observational in nature while the computer experiment studies may require a careful design of experiments. Space-filling designs such as Latin hypercubes (McKay et al., 1979) are the most commonly used in computer experiments. Special classes of designs like nested and sliced designs (Qian, 2009; Qian and Wu, 2009) were developed to meet the needs of specific problems. Finally the interpolating property as described above is desirable for deterministic computer experiments but is not required for spatial statistics.

As the research in computer experiments has rapidly expanded in recent years, it is a timely question to ask if there are new principles that can be used to guide the design and modeling of computer experiments. It is natural to ask if the three principles of Fisher are applicable here. Replication is irrelevant to deterministic experiments because replicating the same input will not produce a different output in deterministic simulation. It is applicable to stochastic experiments but does not rise to the level of a fundamental principle as it is quite straightforward and does not help to address the complexity of the problem. Blocking and randomization are not relevant even to stochastic experiments. How about the three principles for factorial designs discussed in Section 2? Effect sparsity can be invoked if there are many input factors and a majority of them are assumed to be inert. This may hold in the case of sensitivity analysis if the input factors have additive effects on the response (Morris, 1991). Because the unknown function g in computer experiments can be quite complicated, effect hierarchy and heredity principles may be too simplistic to be useful. It is thus a major challenge to formulate new principles that can be used to guide work in computer experiments. One clue may come from how we classify errors in an emulator. According to Haaland and Qian (2011), there are three sources of error: *parameter estimation error, numeric error, and nominal error.* The last two

types are not present in physical experiments as they arise in the computational side of the problem. If we recall that Fisher's three principles are used to control or reduce various sources of errors in physical experiments, some new principles may be formulated for assessing and reducing these three types of errors. Such work will be interesting and novel, especially work on numeric and nominal errors.

A potential difficulty with the kriging technique lies in the computation of the inverse matrix $\mathbf{R}^{-1}$ in (5.2). For large sample size n and/or large input dimension p, the matrix $\mathbf{R}$ can be nearly singular. The numerical instability in computing $\mathbf{R}^{-1}$ can be serious because it can lead to large variability and poor performance of the predictor $\hat{y}$ in (5.2). See Peng and Wu (2014) for some relevant algebraic results. The practical use of kriging for large and/or complex problems has been hindered by this limitation. Various attempts along the stochastic lines to circumvent this problem are available but none can claim to have resolved the difficulty for large practical problems. The main challenge lies in making the trade-off between achieving numerical stability and obtaining high prediction accuracy. An alternative approach is to use some fast and stable numerical approximation to the unknown function $g(\mathbf{x})$ while maintaining a tight bound on the interpolating errors. Methods known to statisticians includes radial basis interpolating functions, smoothing splines, MARS, and neural networks (Hastie et al., 2011). However, some of them do not have a built-in stochastic element to form the basis for performing uncertainty quantification (UQ). UQ is a term coined in applied mathematics. The forward version of UQ is on quantifying the uncertainties in system outputs propagated from uncertain inputs. General information on UQ can be found in the book by Le Maître and Kino (2010). Computer experiments which were developed by statisticians in parallel to work in applied mathematics can be viewed as UQ because any statistical modeling has the necessary stochastic elements to give it the inferential capability, i.e., the capability

to assess stochastic errors in prediction, estimation, testing and variable selection. In the applied mathematics literature, a popular approximation method that possesses the inferential capability is the technique of generalized polynomial chaos (gPC). But this method has some limitations too, especially for high dimensional problems. See Ghanem and Spanos (1991) and Xiu (2010) for details. One challenge for the numerical approach is to develop methods that can perform fast and stable computation, handle large problems (i.e., large sample and dimension), and have inferential capability. Existing theoretical results on the approximation accuracy of these methods often depend on unknown parameters or upper bounds with unknown parameters, and are thus of limited value in practice. Statistical ideas such as cross-validation can be used to derive data-driven bounds to overcome this difficulty. See, for example, Zhang and Qian (2013).

6 Concluding remarks

This paper gives a personalized glimpse of some advances in physical experiments in the last 60 years. Three topics are discussed: effect principles for factorial designs, a new method called CME for de-aliasing aliased effects, and robust parameter design. As in the case of response surface methodology, they were developed primarily in response to needs in industrial experimentation. Although some of the concepts and details in these works are different from those in the Fisherian approach, they clearly show that Fisher's legacy and influence have continued from agricultural experiments to industrial experiments and beyond.

For virtual or computer experiments, Fisher's influence will not diminish but there are more challenges. As argued in Section 5, there is a need of new principles that can guide the design and analysis of computer experiments. For the kriging approach,

the need to find an efficient and stable computational method for large problems is a major challenge. For the numerical approach, the challenge is of a different nature, namely, how to bring in stochastic elements in order to give it the inferential capability? Interaction with the applied mathematics community on the emerging field of uncertainty quantification (UQ) is a promising new opportunity for statisticians. If the interface with applied mathematics can lead to major new work and paradigm shift, it will open a new chapter in the history of DOE.

References

[1] Box, G. E. P. (2013), *An Accidental Statistician: The Life and Memories of George E. P. Box*, New York: Wiley.

[2] Box, G. E. P., and Hunter, J. S. (1961), "The 2^{k-p} Fractional Factorial Designs Part I," *Technometrics*, 3, 311–351.

[3] Box, G. E. P., and Meyer, R. D. (1986), "An Analysis for Unreplicated Fractional Factorials," *Technometrics*, 28, 11–18.

[4] Box, J. F. (1978), *R. A. Fisher: The Life of a Scientist*, New York: Wiley.

[5] Brajac, M., and Morey, C. (1987), "GM10 Chassis Body Marriage Simulation: a Designed Experiment Approach," *Journal of Designed Experiment Case Studies*, 1, 1, 4–31, GM of Canada.

[6] Breiman, L. (1995), "Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, 37, 373–384.

[7] Chipman, H. (1996), "Bayesian Variable Selection with Related Predictors," *Canadian Journal of Statistics*, 24, 17–36.

[8] Cressie, N. (1990), "The origins of kriging," *Mathematical Geology*, 22, 239–252.

[9] Efron, B., Johnstone, I., Hastie, T., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499.

[10] Finney, D. J. (1945), "The Fractional Replication of Factorial Arrangements," *Annals of Eugenics*, 12, 291–302.

[11] Ghanem, R. G., and Spanos, P. D. (1991), *Stochastic Finite Elements: A Spectral Approach*, New York: Springer.

[12] Haaland, B., and Qian, P. Z. G. (2011), "Accurate Emulators for Large-scale Computer Experiments," *The Annals of Statistics*, 39, 2974–3002.

[13] Hamada, M. S., and Wu, C. F. J. (1992), "Analysis of Designed Experiments with Complex Aliasing," *Journal of Quality Technology*, 24, 130–137.

[14] Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2011), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, (2nd ed.), New York: Springer.

[15] Ishikawa, K. (1971), *Guide to Quality Control*, Asia Productivity Organization.

[16] Juran, J. M. (1974), *Quality Control Handbook*, (3rd ed.), New York: McGraw Hill.

[17] Kennedy, M. C., and O'Hagan, A. (2001), "Bayesian Calibration of Computer Models," *Journal of the Royal Statistical Society*, Ser. B, 63, 425–464.

[18] Le Maître O.P., and Kino O.M. (2010), *Spectral Methods for Uncertainty Quantification, With Applications to Fluid Dynamics*, Springer.

[19] Lynch, M., and Walsh, B. (1998), *Genetics and Analysis of Quantitative Traits*, Sinauer Sunderland.

[20] McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, (2nd ed.), London: Chapman and Hall.

[21] McKay, M. D., Beckman, R. J., and Conover, W. J. (1979), "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," *Technometrics*, 21, 239–245.

[22] Montgomery, D. C. (2005), *Design and Analysis of Experiments*, (6th ed.), New York: Wiley.

[23] Morris, M. D. (1991), "Factorial Sampling Plans for Preliminary Computational Experiments," *Technometrics*, 33, 161–174.

[24] Mukerjee, R., and Wu, C. F. J. (2006), *A Modern Theory of Factorial Design*, New York: Springer.

[25] Nelder, J. A. (1977), "A Reformulation of Linear Models (with discussion)," *Journal of the Royal Statistical Society*, Ser. A, 140, 48–77.

[26] Peng, C., and Wu, C. F. J. (2014), "On the Choice of Nugget in Kriging Modeling for Deterministic Computer Experiments," *Journal of Computational and Graphical Statistics*, to appear.

[27] Qian, P. Z. G. (2009), "Nested Latin Hypercube Designs," *Biometrika*, 96, 957–970.

[28] Qian, P. Z. G., Seepersad, C., Joseph, V. R., Allen, J., and Wu, C. F. J. (2006), "Building Surrogate Models Based on Detailed and Approximate Simulations," *ASME Journal of Mechanical Design*, 128, 668–677.

[29] Qian, P. Z. G., and Wu, C. F. J. (2009), "Sliced Space-Filling Designs," *Biometrika*, 96, 945–956.

[30] Santner, T. J., Williams, B. J., and Notz, W. I. (2003), *The Design and Analysis of Computer Experiments*, New York: Springer.

[31] Su, H. and Wu, C. F. J. (2014), "CME Analysis: a New Method for Unraveling Aliased Effects in Fractional Factorial Experiments," manuscript.

[32] Taguchi, G. (1986), *Introduction to Quality Engineering*, Asian Productivity Organization, (Available from UNIPUB, New York).

[33] Taguchi, G. (1987), *System of Experimental Design: Engineering Methods to Optimize Quality and Minimize Costs*, UNIPUB/Kraus International Publications (White Plains) and American Supplier Institute, Inc. (Dearborn).

[34] Wu, C. F. J., and Hamada, M. S. (2000), *Experiments: Planning, Analysis and Parameter Design Optimization*, New York: Wiley.

[35] Wu, C. F. J., and Hamada, M. S. (2009), *Experiments: Planning, Analysis, and Optimization*, (2nd ed.), New York: Wiley.

[36] Wu, C. F. J., and Zhu, Y. (2003), "Optimal Selection of Single Arrays for Parameter Design Experiments," *Statistica Sinica*, 13, 1179–1199.

[37] Xiu, D. (2010), *Numerical Methods for Stochastic Computations: a Spectral Method Approach*, Princeton University Press.

[38] Yates, F. (1935), "Complex Experiments," *Supplement to the Journal of the Royal Statistical Society*, 2, 181–247.

[39] Yates, F. (1937), "The Design and Analysis of Factorial Experiments," *Commonwealth Bureau of Soil Science, Technical Communication*, 35.

[40] Yuan, M., Joseph, V. R., and Lin, Y. (2007), "An Efficient Variable Selection Approach for Analyzing Designed Experiments," *Technometrics*, 49, 430–439.

[41] Yuan, M., Joseph, V. R., and Zou, H. (2009), "Structured Variable Selection and Estimation," *Annals of Applied Statistics*, 3, 1738–1757.

[42] Zhang, Q., and Qian, P. Z. G. (2013), "Designs for Crossvalidating Approximation Models," *Biometrika*, 100, 997–1004.