A Bayesian Variable-Selection Approach for Analyzing Designed Experiments With Complex Aliasing

H. Chipman
Department of Statistics and Actuarial Science
University of Waterloo
Waterloo, Ontario N2L 3G1
Canada
(hachipman@uwaterloo.ca)

M. Hamada
Department of Statistics
University of Michigan
Ann Arbor, MI 48109-1027
(mshamada@stat.lsa.umich.edu)

C. F. J. Wu
Departments of Statistics
and of Industrial and Operations Engineering
University of Michigan
Ann Arbor, MI 48109-1027
(jeffwu@stat.lsa.umich.edu)

Experiments using designs with complex aliasing patterns are often performed—for example, two-level nongeometric Plackett–Burman designs, multilevel and mixed-level fractional factorial designs, two-level fractional factorial designs with hard-to-control factors, and supersaturated designs. Hamada and Wu proposed an iterative guided stepwise regression strategy for analyzing the data from such designs that allows entertainment of interactions. Their strategy provides a restricted search in a rather large model space, however. This article provides an efficient methodology based on a Bayesian variable-selection algorithm for searching the model space more thoroughly. We show how the use of hierarchical priors provides a flexible and powerful way to focus the search on a reasonable class of models. The proposed methodology is demonstrated with four examples, three of which come from actual industrial experiments.

KEY WORDS: Gibbs sampler; Hard-to-control factors; Interactions; Partial aliasing; Plackett–Burman designs; Supersaturated designs.

Nongeometric Plackett–Burman (1946) (PB) designs (i.e., those whose run sizes are not a power of two), such as those with 12, 20, and 24 runs, are popular for screening a large number of two-level factors because of their run-size economy. The analysis of these PB designs has traditionally been confined to main effects only under the assumption that the interactions are negligible. This focus on main effects is due to the complex aliasing patterns of these designs. Consider the 12-run PB design with 11 factors in Table 1, Section 1: For each factor, say X, the main effect is partially aliased with the 45 two-factor interactions not involving X. Because of such complex aliasing patterns, Daniel (1976, p. 294) had reservations about using PB designs even for screening and referred to their complex aliasing patterns as "hazards."

Hamada and Wu (1992) went beyond the traditional approach by showing that interactions could be identified and estimated with reasonable precision from such designs with complex aliasing. They proposed an iterative analysis strategy based on the precepts of *effect sparsity* (i.e., experimental variation attributed to only a few effects) and *effect heredity* (i.e., a significant two-factor interaction occurring with at least one of the corresponding main effects being significant) that exploited the designs' complex aliasing patterns. Recognizing the potential for entertaining interactions, the "hazards" of the nongeometric PB designs could now be viewed as "advantages." For example, in geometric PB designs (i.e., 2^{k-p} fractional factorials), a main effect is either orthogonal to or completely aliased with an interaction so that, if main effect A is completely aliased with interaction BC, the geometric PB design would provide no information about their separate effects; in contrast, for a nongeometric PB design, the two effects are partially aliased and can be decoupled under certain assumptions.

Designs and data with complex aliasing patterns arise in several situations:

1. *Two-level factors.* In addition to the nongeometric 12, 20, 24, and 28 PB designs (Plackett and Burman 1946), Hall (1961) gave four nongeometric 16-run designs.

2. *Multilevel and mixed-level fractional factorials.* 3^{k-p} fractional factorials are examples of multilevel designs in

which complex aliasing arises if each main effect is decomposed into the linear and quadratic contrasts and each two-factor interaction into linear × linear, linear × quadratic, quadratic × linear, and quadratic × quadratic contrasts. $L_{18}(2 \times 3^7)$ and $L_{36}(2^{11} \times 3^{12})$ are examples of mixed-level designs that accommodate both two-level and three-level factors. See Wang and Wu (1991) for many classes of mixed-level designs. Wang and Wu (1992) also considered "nearly orthogonal" designs whose main effects are either orthogonal or nearly orthogonal and that also have complex aliasing patterns.

3. *Hard-to-control factors.* There may be some difficulty in controlling the experimental factors exactly so that the experimental design is not carried out as planned. Consequently, even a 2^{k-p} fractional factorial design will no longer be orthogonal when improperly implemented and therefore will have complex aliasing patterns. Moreover, a mistake may be made in setting the factors levels for a particular run, which will have the same adverse effect.

4. *Supersaturated designs.* Supersaturated designs allow the study of more factors than runs. See recent work by Lin (1993) and Wu (1993) that presented designs that have complex aliasing. In fact, those given by Wu (1993) use the partially aliased interaction columns of the PB designs to accommodate the additional factors.

The analysis strategy of Hamada and Wu (1992) was motivated by the potential infeasibility of performing all-subsets regression with main effects and all two-factor interactions; examples of this include (a) more effects than runs (or observations), (b) computational infeasibility, say with 66 effects for a 12-run PB design, and (c) potential unreasonable models with two-factor interactions and no main effects. The Hamada and Wu (1992) analysis strategy used an iterative stepwise regression approach that addressed (a) and (b) and was guided by the principle of *effect heredity,* which addressed (c). Their strategy did not explicitly impose effect heredity, so models with two-factor interactions without corresponding main effects may still be obtained.

Although providing a feasible alternative to an all-subsets regression, the Hamada and Wu (1992, p. 132) strategy and a modified form (p. 136) explored a small part of the entire model space. More comprehensive searches are needed. Complex aliasing and designs with more effects than runs mean that the model space is very large and may contain different models that explain the available data. The stepwise strategy tends to identify a single model, however. We present a feasible and more comprehensive search that addresses (a)–(c). This Bayesian approach combines the stochastic search variable selection (SSVS) algorithm of George and McCulloch (1993) with priors for related predictors given by Chipman (1996). A suitable class of hierarchical prior distributions focuses the search on a reasonable class of models as suggested by (c) (i.e., that obey effect heredity). The stochastic nature of the search means that all models have positive probability of being visited. In practice, when the data suggest multiple models, the procedure is able to identify them. The stochastic search is "data-guided," so when data suggest that a small subset of models are most likely, reasonable estimates of the probability of these models are available based on many fewer posterior samples than the total number of models.

The article is organized as follows. In Section 1, four examples (three with real data) are given that illustrate the situations in which complex aliasing arises. In Section 2, a Bayesian variable-selection algorithm that incorporates the hierarchical model requirements—that is, *Bayesian hierarchical model selection*—is presented. The experiments given in Section 1 are analyzed in Section 3 using the Bayesian hierarchical model-selection methodology. The article concludes with a discussion in Section 4.

1. EXAMPLES

In this section, examples of four experiments illustrating situations in which complex aliasing arises are given—a screening experiment using a PB 12-run design, a mixed-level design, an experiment with hard-to-control factors, and a supersaturated design.

1.1 Screening Experiment

Table 1 presents a 12-run PB design and illustrates its use in a screening context that can accommodate up to 11 factors labeled A–K. The data were originally constructed by Hamada and Wu (1992) based on the true model $Y = A + 2AB + 2AC + \varepsilon$ with $\varepsilon \sim \mathrm{N}(0, \sigma = .25)$; that is, factors $A, B,$ and C are active with the remaining factors D–K inactive. For an actual experiment that reanalyzed a 12-run PB design to improve the reliability of weld-repaired casts, originally due to Hunter, Hodi, and Eager (1982), see Hamada and Wu (1992).

1.2 Blood-Glucose Experiment Using Mixed-Level Design

Henkin (1986) used an 18-run mixed-level design to study the effect of 1 two-level factor and 7 three-level factors on blood-glucose readings made by a clinical laboratory testing device. Note that all the factors were quantitative. Here we consider only one aspect of the study, which was to identify factors that affect the mean reading. The design and response data are given in Table 2, and the factor names and levels are given in Table 3.

Table 1. Screening Experiment With Plackett–Burman 12-Run Design and Response Data

Design											
A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*Response*
+	+	−	+	+	+	−	−	−	+	−	1.058
+	−	+	+	+	−	−	−	+	−	+	1.004
−	+	+	+	−	−	−	+	−	+	+	−5.200
+	+	+	−	−	−	+	−	+	+	−	5.320
+	+	−	−	−	+	−	+	+	−	+	1.022
+	−	−	−	+	−	+	+	−	+	+	−2.471
−	−	−	+	−	+	+	−	+	+	+	2.809
−	−	+	−	+	+	−	+	+	+	−	−1.272
−	+	−	+	+	−	+	+	+	−	−	−.955
+	−	+	+	−	+	+	+	−	−	−	.644
−	+	+	−	+	+	+	−	−	−	+	−5.025
−	−	−	−	−	−	−	−	−	−	−	3.060

Table 2. Blood-Glucose Experiment With Mixed-Level Design and Response Data

Design								Mean reading
A	*G*	*B*	*C*	*D*	*E*	*F*	*H*	
1	1	1	1	1	1	1	1	97.94
1	1	2	2	2	2	2	2	83.40
1	1	3	3	3	3	3	3	95.88
1	2	1	1	2	2	3	3	88.86
1	2	2	2	3	3	1	1	106.58
1	2	3	3	1	1	2	2	89.57
1	3	1	2	1	3	2	3	91.98
1	3	2	3	2	1	3	1	98.41
1	3	3	1	3	2	1	2	87.56
2	1	1	3	3	2	2	1	88.11
2	1	2	1	1	3	3	2	83.81
2	1	3	2	2	1	1	3	98.27
2	2	1	2	3	1	3	2	115.52
2	2	2	3	1	2	1	3	94.89
2	2	3	1	2	3	2	1	94.70
2	3	1	3	2	3	1	2	121.62
2	3	2	1	3	1	2	3	93.86
2	3	3	2	1	2	3	1	96.10

1.3 Experiment With Hard-to-Control Factors

The design given in Table 4 was used in a real experiment on a wood-pulp production process that studied 11 factors. Quality characteristics such as yield, burst index, and opacity were observed. The process consisted of chemical and mechanical treatments; factors A–G involve the chemical treatment, and factors H–K involve the mechanical treatment. The planned experiment was a PB 20-run design with a centerpoint replicated twice (i.e., the total run size was 22). Data from only 19 runs were available because difficulties were encountered in performing three of the runs from the PB design portion. Notice also that several of the factors were hard to control, notably factors E, I, and K (wood-to-liquid ratio, slurry concentrations at two stages); the planned levels were ±1 in runs 1–17 and 0 in runs 18–19. The actual factor levels and the observed quality characteristic, burst index, are given in Table 4.

1.4 Experiment With Supersaturated Design

Lin (1993) showed that a half-fraction of a PB design could be used as a supersaturated design. He illustrated this with a 28-run PB design with 24 factors from an experiment used to develop an epoxide adhesive system as reported by Williams (1968). The half fraction (based on an unused orthogonal column, yielding runs 1, 3, 4, 6, 8–10, 13, 17, 22–25, 28) of the original design along with the corresponding strip adhesion response data are displayed in Table 5. This illustrates the use of a 14-run design to study 23 factors; note that, in the half fraction, factors 13 and 16 were assigned to the same column so that only factor 13 is reported here.

Table 3. Factor Names and Levels, Blood-Glucose Experiment

Code	Variable	Levels
A	Wash	yes, no
B	Volume in microvial	2.0, 2.5, 3.0 ml
C	Water level in caras	20.0, 28.0, 35.0 ml
D	RMP of centrifuge	2,100, 2,300, 2,500
E	Time in centrifuge	1.75, 3.00, 4.50 minutes
F	Sensitivity absorption	.10–2.5, .25–2.0, .50–1.5
G	Temperature	25, 37, 30°C
H	Dilution	1:51, 1:101, 1:151

2. STOCHASTIC VARIABLE SELECTION

This section reviews one algorithm for variable selection based on the Gibbs sampler [see Smith and Roberts (1993) and references therein for an overview]. The criterion of interest is taken to be the posterior probability of a model conditional on the data that can be obtained using the stochastic search variable-selection (SSVS) algorithm of George and McCulloch (1993). The approach can be outlined as follows for the simplest case of linear regression with normal errors:

$$Y = X'\beta + \sigma\varepsilon, \qquad \varepsilon \sim \mathrm{N}(0,1), \tag{1}$$

where β may contain main effects, interaction effects, or polynomial effects. Importance of effects is captured via an unobserved vector $\boldsymbol{\delta}$ of zeros and ones of length p, the same length as $\boldsymbol{\beta}$. When $\delta_i = 0$, the magnitude of β_i is small and the corresponding predictor is "inactive." When $\delta_i = 1$, the magnitude of β_i is large and the predictor is "active." A normal mixture prior for the coefficients $\boldsymbol{\beta}$ specifies the magnitude of active and inactive effects:

$$f(\beta_i|\delta_i) = \begin{cases} \mathrm{N}(0, \tau_i^2) & \text{if } \delta_i = 0 \\ \mathrm{N}(0, (c_i\tau_i)^2) & \text{if } \delta_i = 1. \end{cases} \tag{2}$$

When $\delta_i = 0, \beta_i$ is tightly centered on 0 and will not have a large effect. The much larger variance ($c_i \gg 1$) when $\delta_i = 1$ allows the possibility of a variable having a large influence. The parameters τ_i and c_i are chosen to represent, respectively, a "small" effect and how many times larger a "large" effect should be.

A prior on $\boldsymbol{\delta}$ corresponds to a prior on the model. The commonly used independence prior implies that the importance of any variable is independent of the importance of any other variable. This is not the case here because the importance of interactions can be assumed to depend on the importance of their corresponding main effects. Hierarchical priors for interactions and polynomial terms, developed by Chipman (1996), are used to formally express these relations in a flexible fashion. These priors are described in Section 2.1.

A prior must also be specified for σ; following George and McCulloch (1993), we take $\sigma^2 \sim \mathrm{IG}(\nu/2, \nu\lambda/2)$, where IG denotes an inverted gamma distribution. This is equivalent to $\nu\lambda/\sigma^2 \sim \chi^2_\nu$.

This specific parameterization is chosen so that a Gibbs sampling approach may be used to obtain the posterior for $\boldsymbol{\delta}$. The Gibbs sampler uses conditional distributions to produce a sequence of samples from the posterior distribution. Repeated draws are made from the conditional distribution of each parameter, conditional on the data and most recently sampled values of the other parameters. The resultant sample is an approximate sample from the joint posterior of the parameters. Such a technique is useful when the posteriors are not available in closed form, which is the case

Table 4. Experiment with Hard-to-Control Factors, Design, and Response Data

Design											
A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*Response*
1	−1.00	1	1	−0.33	−1	1	1	0.74	1	−.89	1.61
−1	−1.00	1	1	1.63	1	−1	1	−1.02	1	−.76	1.97
−1	0.99	−1	−1	−1.04	−1	1	1	−0.55	1	1.85	1.48
1	1.00	−1	1	1.82	−1	−1	1	0.35	1	1.03	.55
1	1.17	−1	−1	0.31	1	1	1	−0.67	1	−1.08	.55
−1	−1.00	−1	−1	1.00	1	−1	1	0.75	−1	−.87	1.59
1	1.00	−1	1	−0.57	1	−1	−1	−1.19	−1	2.26	1.64
−1	−1.00	−1	−1	−0.32	−1	−1	−1	−1.16	−1	−.79	1.50
1	1.00	1	−1	1.69	−1	1	−1	−1.20	−1	−.87	1.97
−1	−1.00	−1	1	1.32	−1	1	1	−1.17	−1	2.17	1.67
1	−0.98	1	−1	1.57	−1	−1	−1	−1.41	1	1.12	1.52
−1	1.00	1	1	1.61	−1	1	−1	−0.77	−1	−.40	4.37
−1	1.00	1	−1	−1.06	1	1	1	0.45	−1	2.32	2.38
1	1.00	1	1	−0.76	1	−1	1	−0.62	−1	−.83	2.04
−1	−1.00	1	1	−0.33	1	1	−1	−1.69	1	−1.38	2.24
1	−1.00	−1	1	1.36	1	1	−1	3.35	−1	.66	1.76
−1	1.00	−1	1	−0.23	−1	−1	−1	1.45	1	−.65	1.73
0	0.00	0	0	−0.10	0	0	0	−0.09	0	.39	1.74
0	0.00	0	0	0.05	0	0	0	0.58	0	.16	1.76

here. Discussion of this technique in general was given by Smith and Roberts (1993); George and McCulloch (1993) and Chipman (1996) discussed its application to variable selection. Here, as given by George and McCulloch (1993), the algorithm consists of a multivariate normal draw for $\boldsymbol{\beta}|\boldsymbol{\delta}, \sigma$, an inverse gamma draw for $\sigma|\boldsymbol{\beta}, \boldsymbol{\delta}$, and p Bernoulli draws for $\delta_i|\boldsymbol{\beta}, \sigma, \{\delta_j\}_{j\neq i}$. One implementation issue is how many starting points to use and how long to run each chain. In the cases explored here, correlations between samples as far apart as lag 20 may exist, so it is most efficient to run one long chain (typically 50,000 draws) and store every kth (typically 5th) draw. Duplicate runs may be used to assess the accuracy of the results and determine how long the chain should be run.

In variable-selection problems with many effects and large correlations, the "mixing" behavior of the chain is of interest. When started from different points, will the chain reach the same models? Our experience is positive. In the glucose example, the time to reach the most probable model under weak heredity (BH^2, B^2H^2) was recorded using four start points. The start points were null, linear only, everything except linear, and all-term models. Ten chains were started from each point. The median time to reach the indicated model was under 150 steps in all four cases, and the maximum ranged from 345 to 793 steps. This represents a small fraction of the 50,000 steps used in the full example. There were 113 effects with correlations in the range of (−.68, .67). Cowles and Carlin (1996) provided an overview of convergence and mixing diagnostics for Markov-chain Monte Carlo.

2.1 Hierarchical Priors for Variable Selection

A prior for $\boldsymbol{\delta}$ should capture the dependence relation between the importance of a higher-order term and those lower-order terms from which it was formed. Consider a simple example in which there are three main effects A, B, C and three two-factor interactions AB, AC, and BC. The importance of, say, AB will depend on whether the main effects A and B are included in the model. If neither are, then the interaction seems less plausible and more difficult to explain. This belief can be expressed in the prior

Table 5. Experiment With Supersaturated Design and Response Data

Design																							
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*	*Response*
1	1	1	−1	−1	−1	1	1	1	1	1	−1	1	−1	−1	1	−1	−1	1	−1	−1	−1	1	133
1	−1	−1	−1	−1	−1	1	1	1	−1	−1	−1	1	1	1	−1	1	−1	−1	1	1	−1	−1	62
1	1	−1	1	1	−1	−1	−1	−1	1	−1	1	1	1	1	1	−1	−1	−1	−1	1	1	−1	45
1	1	−1	1	−1	1	−1	−1	−1	1	1	−1	1	−1	1	−1	1	1	1	−1	−1	−1	−1	52
−1	−1	1	1	1	1	−1	1	1	−1	−1	−1	1	−1	1	1	−1	−1	1	−1	1	1	1	56
−1	−1	1	1	1	1	1	−1	1	1	1	−1	−1	1	1	1	1	1	1	1	1	−1	−1	47
−1	−1	−1	−1	1	−1	−1	1	−1	1	−1	1	1	1	−1	1	1	1	1	1	−1	−1	1	88
−1	1	1	−1	−1	1	−1	1	−1	1	−1	−1	−1	−1	−1	−1	−1	1	−1	1	1	1	−1	193
−1	−1	−1	−1	−1	1	1	−1	−1	−1	1	1	−1	−1	1	1	1	−1	−1	−1	−1	1	1	32
1	1	1	1	−1	1	1	1	−1	−1	−1	1	−1	1	1	1	−1	1	−1	1	−1	−1	1	53
−1	1	−1	1	1	−1	−1	1	1	−1	1	−1	−1	1	−1	−1	1	1	−1	−1	−1	1	1	276
1	−1	−1	−1	1	1	1	−1	1	1	1	1	1	−1	−1	−1	−1	1	−1	1	1	1	1	145
1	1	1	1	1	−1	1	−1	1	−1	−1	1	−1	−1	−1	−1	1	−1	1	1	−1	1	−1	130
−1	−1	1	−1	−1	−1	−1	−1	−1	−1	1	1	−1	1	−1	−1	−1	−1	1	−1	1	−1	−1	127

for $\boldsymbol{\delta} = (\delta_A, \ldots, \delta_{BC})$ as follows:

$$\Pr(\boldsymbol{\delta}) = \Pr(\delta_A)\Pr(\delta_B)\Pr(\delta_C)\Pr(\delta_{AB}|\delta_A,\delta_B) \times \Pr(\delta_{AC}|\delta_A,\delta_C)\Pr(\delta_{BC}|\delta_B,\delta_C). \quad (3)$$

In (3), two principles are used to obtain the simplified form. First, the *conditional independence principle* assumes that, conditional on first-order terms, the second-order terms $(\delta_{AB}, \delta_{AC}, \delta_{BC})$ are independent. Independence is also assumed for main effects. The *inheritance principle* assumes that the importance of a term depends only on those terms from which it was formed, implying that $\Pr(\delta_{AB}|\delta_A,\delta_B,\delta_C) = \Pr(\delta_{AB}|\delta_A,\delta_B)$.

The exact nature of this dependence on "parent" terms is defined by the components of the joint probability in (3). For example, the probability that the term AB is active $\Pr(\delta_{AB} = 1|\delta_A,\delta_B)$ takes on four different values:

$$\Pr(\delta_{AB} = 1|\delta_A,\delta_B) = \begin{cases} p_{00} & \text{if } (\delta_A,\delta_B) = (0,0) \\ p_{01} & \text{if } (\delta_A,\delta_B) = (0,1) \\ p_{10} & \text{if } (\delta_A,\delta_B) = (1,0) \\ p_{11} & \text{if } (\delta_A,\delta_B) = (1,1). \end{cases} \quad (4)$$

Here, we choose p_{00} small (e.g., .01), p_{01} and p_{10} larger (e.g., .10), and p_{11} largest (e.g., .25). This represents the belief that a two-factor interaction without parents is quite unlikely, one with a single parent is more likely, and one with both parents is most likely. The term *relaxed weak heredity* will refer to this prior, and setting $p_{00} = 0$ yields *strict weak heredity*. Probabilities of less than .5 for both main effects and interactions represent the belief that relatively few terms are active. Such *effect sparsity* is a reasonable assumption for screening experiments.

This prior may be generalized to polynomials and interactions involving polynomials. Consider a simple example, with fourth-order term A^2B^2; third-order terms AB^2, A^2B; second-order terms A^2, AB, B^2; and first-order terms A and B. We consider the parents of a term to be those terms of the next smallest order that can form the original term when multiplied by a main effect. Here, A^2B^2 has parents A^2B (because multiplication by B produces A^2B^2) and AB^2. Some terms (such as A^2) will have only one parent (e.g., A). We assume that the importance of a term depends only on its parents, an assumption called the *immediate inheritance principle*. The conditional independence principle (which now says that terms of a given order are independent given all lower-order terms) is applied as before. See Chipman (1996) for details and a discussion of other forms of hierarchical dependence between variables.

Another interesting class of predictors is qualitative predictors such as treatment, supplier, or location. Although not present in our examples, such variables arise in screening experiments, often in the form of three-level factors. Dummy variables are commonly used in such a situation, but typically one wants either all or none of the variables to be included in the model. Chipman (1996) gave a prior for $\boldsymbol{\delta}$ that forces all associated dummy variables to be active or inactive. Extensions to interactions involving qualitative factors are straightforward.

2.2 Choice of Additional Prior Parameters

Several prior parameters need to be specified. The normal mixture prior on $\boldsymbol{\beta}$ has parameters τ and c, and the inverse gamma prior for σ has parameters ν and λ. Because this methodology is used as a tool rather than strictly for Bayesian reasons, we view these parameters as tuning constants as well as representations of prior information.

As Box and Meyer (1986) did, we use $c = 10$, which indicates that an important effect is an order of magnitude larger than an unimportant one. For the choice of τ, we take, as did George and McCulloch (1993),

$$\tau_j = \Delta Y / 3\Delta X_j, \quad (5)$$

where ΔY represents a "small" change in Y and ΔX_j represents a "large" change in X_j. This implies that, if $\delta_j = 0$, even a large change in X_j is unlikely to produce anything more than a small change in Y. $\Delta X_j = \max(X_j) - \min(X_j)$ is used unless the X_j are not set at their intended levels. Details of this latter case are given in Section 3.3. A value of ΔY still must be chosen. When expert opinions are not available,

$$\Delta Y = \sqrt{\text{var}(Y)}/5,$$

where $\text{var}(Y)$ is the sample variance of the response without any regression, is found to work well in practice. This choice corresponds to the belief that, after a model is fit to the data, σ will be roughly 20% of the unadjusted standard deviation.

Choice of τ can be quite influential for the posterior because it defines the magnitude of an important coefficient. A parameter of this nature is usually a component of Bayesian model selection. For example, Box and Meyer (1993) used γ, a prior variance for β that serves the same role. They recommended choosing the posterior mode of this parameter; here we use several values near the first guess to assess sensitivity of posteriors to this parameter.

Sensitivity to τ may be viewed as an advantage of the procedure because it allows the user to choose different models by fine-tuning the parameter τ. Unless there is a strong belief or past knowledge to suggest the choice of τ, the appropriateness of a chosen τ may be judged by the models (or posterior model probabilities) it generates. Too few or too many active terms may be considered inappropriate. Our deliberations in the choice of models in Sections 3.1 and 3.3 illustrate this point well. Often the experimenter has much better prior knowledge about model size than about the numeric value of a prior parameter like τ.

An improper prior (i.e., $\nu = 0$) for σ is not appropriate because this allows σ to be close to 0. When the number of predictors outnumber the observations, this prior will result in a posterior with very small σ values and many terms in the model. An informative prior on σ corrects this. The assumption that $\sigma \approx \sqrt{\text{var}(Y)}/5$ suggests that a prior on σ with a mean equal to $\sqrt{\text{var}(Y)}/5$ be used. Among these priors, the desirable spread may be attained by selecting a prior with an upper quantile (say 99%) that is near $\sqrt{\text{var}(Y)}$. This approach often yields a value of ν near 2, which corresponds to a reasonably uninformative prior. The value of λ

changes from experiment to experiment because it depends on the scale of the response measurement.

3. ANALYSIS

The Bayesian hierarchical model-selection methodology presented in Section 2 will be illustrated using the four examples given in Section 1.

3.1 Screening Design Experiment

The data were originally constructed by Hamada and Wu (1992) to illustrate that their stepwise strategy for variable selection could have difficulties identifying interactions if the corresponding main effects were smaller. The stepwise nature of their procedure caused it to miss all three active terms, which suggests that the proposed approach might be more effective.

Because the true model is known, we used several different priors to assess the influence of different parameters. For the prior on σ, $(\nu, \lambda) = (1.5, .0015), (1.5, .038), (1.5, .15)$ were investigated. The parameter λ was varied because it represents the location of the prior on σ. Choosing $\nu = 1.5$ implies a reasonably uninformative prior. The first and third priors are intended to represent "extreme" situations, whereas the middle value was chosen using the $\sqrt{\text{var}(Y)}$ rule. Because conclusions were insensitive to λ, only results for the automatically chosen $\lambda = .038$ are given here.

We also need to specify the probabilities that factors are active. A relaxed weak heredity prior will be used (see Sec. 2.1), with

$$\Pr(\delta_A = 1) = .25$$

$$\Pr(\delta_{AB} = 1) = \begin{cases} .01 & \text{if } \delta_A = \delta_B = 0 \\ .10 & \text{if } \delta_A \neq \delta_B \\ .25 & \text{if } \delta_A = \delta_B = 1. \end{cases}$$

A, B, and AB represent arbitrary main effects and interactions. This prior allows interactions to be active if only one parent term is active, and even if both parents are inactive, there is a small probability that the interaction will be active.

The posteriors are insensitive to choice of c, but changes in τ have more influence. The estimate based on the rule of thumb (5) is $\tau^* = .105$. To examine the robustness of conclusions to τ, we performed analyses with $\tau^*/2, \tau^*$, and $2\tau^*$.

The Gibbs algorithm was started from a model with no active terms, run 10,000 cycles, and every tenth sample saved. Examination of autocorrelations and output from several independent runs confirmed that convergence and mixing occurred quickly, and that the suggested number of runs is adequate. Table 6 and Figure 1 give joint and marginal posteriors for three different values of τ. Although the correct model has the most mass in all three cases, there is considerable dependence on τ. When a small value is used ($\tau^*/2$), too many effects are considered "large," leading to less model certainty. As τ increases to τ^*, there is less model uncertainty, and the correct model receives the most mass. As τ increases further to $2\tau^*$, there is again less model certainty. The behavior of the algorithm for large τ values is better understood by looking at the marginal distributions. From them, it is clear that no term appears to be important, as one would expect when τ is too large. The correlations between the many candidate terms result in many models receiving some posterior mass, many of them nonsensical.

In this case, the method works quite well and clearly identifies the correct model. It succeeds because it searches

Table 6. Screening Experiment Posterior Model Probabilities (true model A, B, AB)

Model	$\tau^*/2$	τ^*	$2\tau^*$
A, AB, AC	.103	.325	.094
A, C, AB, AC	.016	.039	.009
A, B, AB, AC	.018	.022	.008
I, DF	—	—	.013
K, GJ	—	—	.009

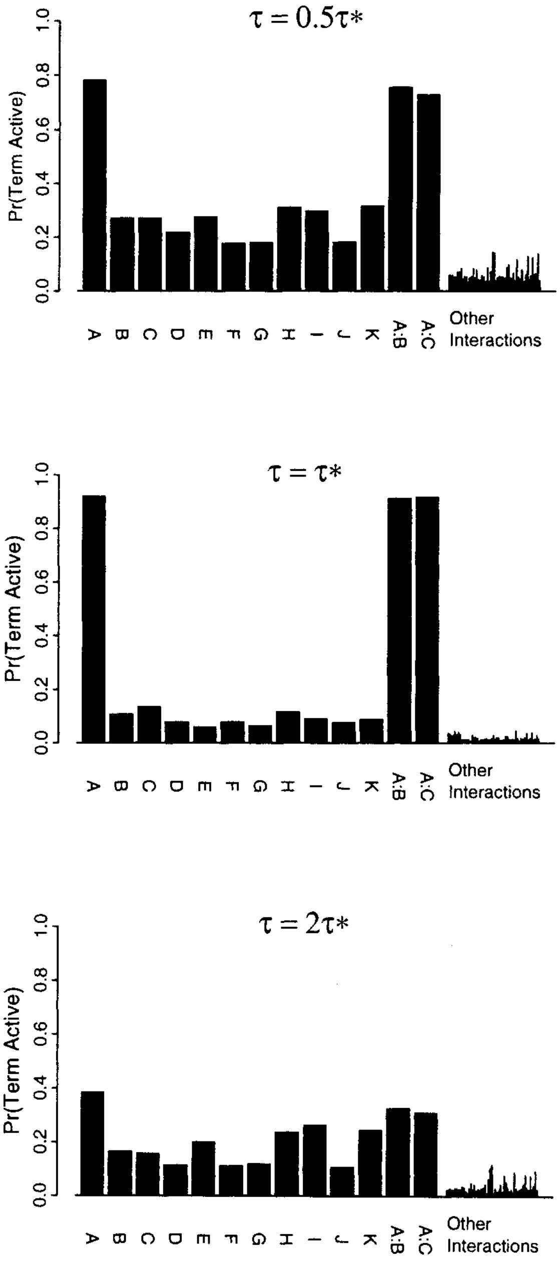

Figure 1. Screening Experiment Marginal Posterior Probabilities. Bars for interactions with small probabilities have narrow widths.

the entire model space, in a nonstepwise fashion. The prior for δ serves to focus attention on certain elements of the model space while not totally excluding others. This class of likely models is much larger than the one used by Hamada and Wu (1992).

3.2 Blood-Glucose Experiment

Recall that the blood-glucose experiment [also analyzed by Hamada and Wu (1992)] consists of continuous factors in either two (factor A) or three levels. The three-level factors are quantitative and allow entertainment of polynomial terms and interactions. There are 15 df for the main effects and an additional 98 two-factor interactions to be considered. None are totally correlated so that all 113 candidate variables will be considered simultaneously. Linear and quadratic terms will be used throughout, with interactions having four components—linear by linear, linear by quadratic, quadratic by linear, and quadratic by quadratic. Choice of $(\nu, \lambda) = (2, 1.289)$ is based on the automatic procedure of Section 2.2. Values of $c = 10$ and $\tau = 2\tau^*$, where τ^* is the automatic choice, were used. The hierarchical priors used are

$$\begin{aligned}
\Pr(\delta_A = 1) &= .25 \\
\Pr(\delta_{A^2} = 1|\delta_A) &= \begin{cases} .01 & \text{if } \delta_A = 0 \\ .25 & \text{if } \delta_A = 1 \end{cases} \\
\Pr(\delta_{AB^2} = 1|\delta_{AB}, \delta_{B^2}) &= \begin{cases} .01 & \text{if } \delta_{AB} = \delta_{B^2} = 0 \\ .10 & \text{if } \delta_{AB} \neq \delta_{B^2} \\ .25 & \text{if } \delta_{AB} = \delta_{B^2} = 1. \end{cases}
\end{aligned}$$

This is a challenging problem because there are only 18 observations and 113 terms from which to choose. Because there are so many variables, there will likely be many models that fit the data well and probably quite a few parsimonious ones. The hierarchical priors will be useful here because they will focus attention on good models that also make sense.

The complexity of the problem is apparent in the simulation, which takes much longer to mix sufficiently. Every fifth draw from a chain of length 50,000 was found sufficient. When relaxed weak heredity priors are used, the most probable model contains two terms, BH^2 and B^2H^2. This model clearly violates even weak heredity, so there must be a good reason for its large mass. The 10 most probable models are given in Table 7. The prevalence of models that contain both BH^2 and B^2H^2 may suggest a nonlinear relation involving these two variables (B = volume of blood, H = dilution ratio). That none of the models obey weak heredity may be an indication of the need for transformations.

Rerunning the algorithm with strict weak heredity (0 replacing probabilities of .01) gives the results in Table 8. We see that the best model is a superset of the previous best, with the appropriate terms for weak heredity added (namely, B and BH). Other models involving EF also appear possible but less likely. The fit of this model, indicated by an R^2 of .86, is quite good. Both models are improvements over the model Hamada and Wu (1992) reported with E^2, F^2, and EF and $R^2 = .68$, a model that does not obey the current definition of weak heredity. The additional information gained from comparing posteriors originating from weak and strict forms of the prior tells us that it is the higher-order interactions between B and H that really drive the model, which may indicate that caution should be exercised in identifying a single "best" model.

Table 7. Blood-Glucose Experiment Posterior Model Probabilities, Relaxed Weak Heredity Prior

Model	*Prob.*	*R^2*
BH^2, B^2H^2	.183	.7696
B, BH^2, B^2H^2	.080	.8548
B, BH, BH^2, B^2H^2	.015	.8601
F, BH^2, B^2H^2	.014	.7943
GE, BH^2, B^2H^2	.013	.8771
AH^2, BH^2, B^2H^2	.009	.8528
G^2D, BH^2, B^2H^2	.009	.8517
A, BH^2, B^2H^2	.008	.7938
B, B^2, BH^2, B^2H^2	.008	.8864
H, BH^2, B^2H^2	.008	.7855

One of these models was also identified by Jan and Wang (1994). They identified a model with B, BH^2, and B^2H^2 terms, similar to the one identified by our procedure. In fact the same model can be identified using a modified strategy recommended by Hamada and Wu (1992, p. 136). Both the Hamada–Wu and Jan–Wang procedures find only one optimal model rather than a set of plausible ones.

The two model-probability posteriors (under relaxed and strict weak heredity) are quite different. Although the change in priors for δ from probabilities of .01 (relaxed) to 0 (strict) seems small, this represents a substantial change. The strict prior states that numerous models are impossible, whereas the relaxed prior merely downweights them. Some of these downweighted models fit the data well, so there is a large shift in the posterior when the priors are changed.

If hierarchical priors are replaced with an independence prior, the posterior becomes much less peaked. In this problem, if the same priors are used but with interactions and higher-order terms having $\Pr(\delta_i = 1) = .1$ independently of other terms, the most probable model is visited only three times in 10,000, a rough posterior probability of .0003. This illustrates that, without hierarchical priors, only marginal distributions contain relevant information in complex aliasing problems. It is not sufficient to set the independent probabilities small because there are still too many possible models that are nonsensical.

Table 8. Blood-Glucose Experiment Posterior Model Probabilities, Strict Weak Heredity Prior

Model	*Prob.*	*R^2*
B, BH, BH^2, B^2H^2	.146	.8601
B, BH, B^2H, BH^2, B^2H^2	.034	.8828
H, H^2, BH^2, B^2H^2	.033	.7903
H, BH, BH^2, B^2H^2	.031	.7908
F, F^2, DF, D^2F, EF	.024	.8835
H, H^2, AH^2, BH^2, B^2H^2	.017	.8735
B, B^2, BH, BH^2, B^2H^2	.013	.8917
B, H, BH, BH^2, B^2H^2	.013	.8760
B, H, H^2, BH^2, B^2H^2	.008	.8756
E, E^2, CE, EF	.008	.6979

To relate our proposed procedure to more traditional methods, (traditional) stepwise and best subsets regressions were run on the dataset. If a stepwise significance level of .05 is used, a six-term model containing $BH^2, B^2H^2, GE, AH^2, H,$ and EF^2 is identified. The ordering of the terms is the order in which they were entered. Other models identified as promising by our procedure are not found. This is not because these models are much worse but because of the limited nature of the stepwise search. Moreover, none of the models identified by stepwise selection obey weak heredity.

For a more complete search than stepwise, all subsets selection was carried out. To make the search feasible, only models with six or fewer terms were considered. The procedure took three days to run on a SPARCstation 20. The best 50 models of each size were recorded. None of these 300 models obey weak or strong heredity. Although these models have a higher R^2, they are not ones that would be used in practice because they do not obey heredity rules.

3.3 Hard-to-Control-Factors Experiment

As discussed in Section 1.3, some predictors could be controlled during the experiment, resulting in complex aliasing. In this experiment, the variables A–G and H–K are considered to be noninteracting groups so that interactions between them are not entertained. All other interactions are considered. Although all the predictors are continuous, only a single quadratic term may be considered because of the structure of the center run (the $-1, +1$ levels all map to 1, resulting in very high correlations between all quadratic terms). This results in a total of 39 predictor terms.

The prior parameters (ν, λ) are chosen to be (2, .00458), giving a mean of .12 for σ and an upper 99% quantile of .66, slightly larger than the unadjusted standard deviation of Y, calculated to be .60. The ranges of X_j's will not be used to determine ΔX_j here because of the large outliers in some of the uncontrolled predictors. Instead we shall assume that the original settings of $(-1, +1)$ represent large changes and take $\Delta X_j = 2$ for all main effects and interactions except A^2, which will have a value of 1. The $\Delta Y / 3\Delta X_i$ rule is used as a starting point for τ. The posteriors have an average of 10 active terms with this automatic choice of τ. Because more parsimonious models are desired, the procedure was rerun after doubling the value of τ. Every fifth draw from a run of length 50,000 was used.

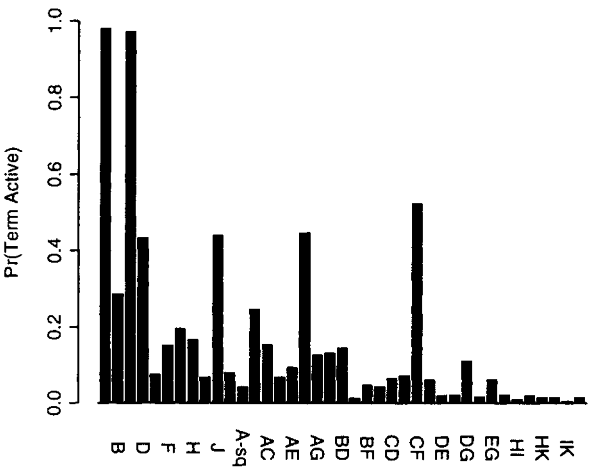

Figure 2. Hard-to-Control-Factors Experiment Marginal Posterior Probabilities. Every alternate bar is labeled. The full set of labels are A. . .K, A^2, and interactions (A. . .G) × (A. . .G), (H. . .K) × (H. . .K).

Posterior probabilities for the model are displayed in Table 9, and marginal posteriors are given in Figure 2. The small probability (.0458) on the most probable model indicates considerable model uncertainty. The terms A, C, AF, CF seem important because they appear in almost all of the most probable models and have high marginal probabilities of being active.

3.4 Supersaturated-Design Experiment

When analyzing the 14-run supersaturated design, we restricted our attention to main-effect-only models because there are already more effects than observations. Because the supersaturated design is a screening design, this seems to be a reasonable approach. The automatic procedures were used to choose the regression coefficient priors, and the hierarchical priors of Section 3.1 were used in the variable selection. The values used were $\lambda = 67, \nu = 2, \tau = 2.31, c = 10$, which appear to produce reasonable results with no modifications. The joint posteriors, which are based on every fifth draw from 50,000 cycles of the chain, are given in Table 10.

The results suggest that factors 4, 12, 15, 20 (and perhaps 10) are active, as Lin (1993) found from his analysis. Although the proposed Bayesian methodology did not find any different results, its more thorough search and flexible priors will work to its advantage in larger experiments that have more candidate models.

Table 9. Hard-to-Control-Factors Experiment Posterior Model Probabilities

Model	Prob.	R^2
A, C, AF, CF	.0458	.7973
A, C, D, AF, CF	.0200	.8488
A, C, F, AF, CF	.0127	.7986
A, C, J, AF, CF	.0123	.8425
A, C, D, J, AF, CF	.0100	.9034
A, B, C, D, J, DG	.0079	.8886
A, C, AC, AF, CF	.0065	.8209
A, C, D, G, J, BD	.0063	.8854
A, C, J, AB	.0059	.7480
A, C, D, J, AB	.0054	.8206

Table 10. Supersaturated-Design Experiment Posterior Model Probabilities

Model	Prob.	R^2
4 12 15 20	.0266	.955
4 10 12 15 20	.0259	.973
4 10 11 12 15 20	.0158	.987
4 12 15 20 21	.0120	.969
4 11 12 15 20	.0082	.966
4 7 10 11 12 15 20	.0076	.998
1 4 12 15 20	:0074	.970
4 12 13 15 20	.0071	.964
1 4 10 12 15 20	.0066	.982
1 4 12 15 17 20	.0064	.978

4. DISCUSSION

Data with complex aliasing arise in numerous situations—experiments using two-level nongeometric screening designs, multilevel and mixed-level fractional factorial designs and nearly orthogonal designs, experiments with hard-to-control factors and with mistakes in setting the factor levels, and experiments using supersaturated designs. Moreover, observational data will typically have complex aliasing. Because data with complex aliasing arise often, an efficient analysis methodology is desirable. Hamada and Wu (1992) showed that information about interactions could be obtained in such situations and proposed an iterative guided stepwise regression strategy.

This article presents a more efficient methodology because it searches the model space more thoroughly, much like an all-subsets regression except that a plausible class of models is considered. Moreover, the proposed methodology requires much less computation because the search is done stochastically rather than fitting all possible models. A Bayesian approach combined with the recent advances in Bayesian computing provides a quick and easy implementation of this strategy. The flexible hierarchical priors of Section 2.1 provide a powerful way to concentrate the search on a reasonable class of models. An advantage of the Bayesian approach is that models strongly suggested by the data but that fall outside the reasonable class of models defined by the hierarchical priors can be identified. Moreover, the proposed methodology can identify several (perhaps incompatible) models that explain the data equally well. We used the proposed methodology on Hamada and Wu's (1992) constructed example 5, a 12-run PB design with true model $Y = 2A + 4C + 2BC - 2CD + \varepsilon$, where ε is normally distributed with mean 0 and standard deviation .5. The proposed methodology quickly found several incompatible but equally plausible models (i.e., models that obey effect heredity). This illustrates a limitation of the PB designs that may require additional experimentation to resolve. Nevertheless, it is important for the experimenter to know that several models fit the data well. Finally, the posterior probability of a model provides a calibrated measure of a model's goodness but does not require adjustment for the number of effects in the model. In fact, the posterior probability uses prior information about relationships between terms to discriminate between models with similar R^2, as seen in Tables 7–10. Note that the best model in the examples not only had a larger posterior probability but, interestingly enough, also had fewer effects than the next best model.

If the search procedure identifies several models with comparable posterior probabilities or when the largest posterior probability is small, it is an indication that the data are not informative enough about the exact model. The investigator may use knowledge in the substantive areas to choose among the models. Otherwise a follow-up experiment may be conducted to resolve model uncertainty. See Meyer, Steinberg, and Box (1996) for a related method of constructing follow-up designs.

Our article adopted the SSVS algorithm of George and McCulloch (1993) to obtain the model posterior probabilities. The key aspects of this method are the stochastic search and hierarchical priors. Our hierarchical priors could be used in conjunction with other stochastic search approaches—for example, those of Raftery, Madigan, and Hoeting (1997) and Geweke (1995) and a conjugate version of SSVS (George and McCulloch 1997). We feel that the choice of algorithm will not influence the conclusions reached because most approaches are similar in spirit. Hierarchical priors could be used with methods that enumerate the entire model space (e.g., Mitchell and Beauchamp 1988; Box and Meyer 1986, 1993). Such methods would be too time consuming in complex aliasing problems, however.

Box and Meyer (1993) proposed an alternative Bayesian approach for analysis of complex aliasing data. They focused on factors rather than specific effects. Their proposal can be summarized as follows. Suppose there are k factors. Using an independence prior on the "factors," each factor has prior probability π of being active. Then, for each of the 2^k subsets of the k factors, a posterior model probability given the data (for a specific model) is calculated. Say, there are i factors in a particular subset. Then the corresponding model has all the main effects and two-factor and three-factor interactions (provided that i is at least 2 and 3, respectively). Note that the number of effects in some of the models will exceed the number of observations. Box and Meyer (1993) used an independence prior for all the effects—that is, regression coefficients β, which does not differentiate between main effects and interactions. The posterior model probabilities are calculated directly, which can be computationally intensive. Active factors are identified using marginal posterior probabilities—that is, the sum of posterior probabilities for all the models given previously containing a particular factor. In contrast, our proposed methodology focuses on effects rather than factors and, in addition to marginal posteriors, considers joint posterior probabilities—namely, posterior probabilities of models. Our methodology requires less computation than an all-subsets approach or exhaustive search (such as that used by Box and Meyer) because the search through the model space is done stochastically. Moreover, the search is focused on, though not restricted to, a class of reasonable models through the specification of flexible hierarchical priors.

ACKNOWLEDGMENTS

We thank the associate editor and two referees for many insightful comments on earlier versions of this article. H. Chipman was supported by funding from the University of Chicago Graduate School of Business. M. Hamada was supported by research grants from General Motors of Canada Limited, the Manufacturing Research Corporation of Ontario, and the Natural Sciences and Engineering Research Council of Canada. C. F. J. Wu was supported by National Science Foundation grant DMS 9404300.

[Received March 1995. Revised March 1997.]

REFERENCES

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

——— (1993), "Finding the Active Factors in Fractionated Screening Experiments," *Journal of Quality Technology*, 25, 94–105.

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

Cowles, M. K., and Carlin, B. P. (1996), "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review," *Journal of the American Statistical Association*, 91, 883–904.

Daniel, C. (1976), *Applications of Statistics to Industrial Experiments*, New York: Wiley.

George, E. I., and McCulloch, R. E. (1993), "Variable Selection Via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

——— (1997), "Approaches to Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373.

Geweke, J. (1995), "Variable Selection and Model Comparison in Regression," in *Proceedings of the Fifth Valencia International Meeting on Bayesian Statistics*, eds. J. O. Berger, J. M. Bernardo, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press.

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

Hall, M., Jr. (1961), "Hadamard Matrices of Order 16," Research Summary 36-10, pp. 1, 21–26, Jet Propulsion Laboratory, Pasadena, CA.

Henkin, E. (1986), "The Reduction of Variability of Blood Glucose Levels," in *Fourth Supplier Symposium on Taguchi Methods*, Dearborn, MI: American Supplier Institute, pp. 758–785.

Hunter, G. B., Hodi, F. S., and Eager, T. W. (1982), "High-Cycle Fatigue of Weld Repaired Cast Ti-6Al-4V," *Metallurgical Transactions*, 13A, 1589–1594.

Jan, H. W., and Wang, P. C. (1994), "Analysis of Experimental Data From Orthogonal Main-Effect Plans," unpublished manuscript.

Lin, D. K. J. (1993), "A New Class of Supersaturated Designs," *Technometrics*, 35, 28–31.

Meyer, R. D., Steinberg, D. M., and Box, G. (1996), "Follow-up Designs to Resolve Confounding in Multifactor Experiments," *Technometrics*, 38, 303–313.

Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1036.

Plackett, R. L., and Burman, J. P. (1946), "The Design of Optimum Multifactorial Experiments," *Biometrika*, 33, 305–325.

Raftery, A., Madigan, D., and Hoeting, J. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179–191.

Smith, A. F. M., and Roberts, G. O. (1993), "Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society*, Ser. B, 55, 3–23.

Wang, J. C., and Wu, C. F. J. (1991), "An Approach to the Construction of Asymmetrical Orthogonal Arrays," *Journal of the American Statistical Association*, 86, 450–456.

——— (1992), "Nearly Orthogonal Arrays With Mixed Levels and Small Runs," *Technometrics*, 34, 409–422.

Williams, K. R. (1968), "Designed Experiments," *Rubber Age*, 100, 65–71.

Wu, C. F. J. (1993), "Construction of Supersaturated Designs Through Partially Aliased Interactions," *Biometrika*, 80, 661–669.