A Two-Stage Bayesian Model Selection Strategy for Supersaturated Designs

Scott D. BEATTIE, Duncan K. H. FONG, and Dennis K. J. LIN

Eli Lilly & Company and The Pennsylvania State University

In early stages of experimentation, one often has many candidate factors of which only few have significant influence on the response. Supersaturated designs can offer important advantages. However, standard regression techniques of fitting a prediction line using all candidate variables fail to analyze data from such designs. Stepwise regression may be used but has drawbacks as reported in the literature. A two-stage Bayesian model selection strategy, able to keep all possible models under consideration while providing a level of robustness akin to Bayesian analyses incorporating noninformative priors, is proposed. The strategy is demonstrated on a well-known dataset and compared to competing methods via simulation.

KEY WORDS: Intrinsic Bayes factor; Markov chain Monte Carlo; Stochastic search variable selection.

1. INTRODUCTION

In early stages of industrial experimentation, one often has a large set of candidate factors believed to have possible significant influence on the response of interest, although it is reasonable to assume that only a small fraction are influential, a condition known as *effect sparsity* (Box and Meyer 1986). Since experiments can be costly, an efficient use of experimental units is the employment of supersaturated designs (Lin 1999), in which the number of observations is less than the number of parameters to be estimated. These designs are useful as screening tools, as the paring of candidate factors is performed in a cost-efficient manner.

The removal of truly influential factors causes obvious problems in follow-up experiments (Lin 1995a, Pan 1999). At the same time, the inclusion of inactive factors can lead to unnecessary costs later (Westfall, Young, and Lin 1998). In the supersaturated design setting, standard regression techniques of fitting a prediction line using all candidate variables fail. The normal equations cannot be solved uniquely, so that some parameters are not estimable. Specialized model selection techniques are needed.

Classical statistical analysis has employed model quality criteria such as p-values, adjusted R^2, s^2, C_p, PRESS, AIC, and BIC within sequential procedures like forward or stepwise selection (Draper and Smith 1998). Inadequacies of classical techniques are the apparent fallibility of p-values (Berger and Sellke 1987), the inflation of Type I errors in sequential procedures (Westfall et al. 1998), as well as the lack of an overall Type I error rate due to the random number of steps taken, and the lack of a probabilistic comparison (for example, p-value) between two competing but nonnested models.

Bayesian methods are able to supplement observational information using prior information on the parameters and allow straightforward computation of posterior probabilities using, for example, Gibbs sampling algorithm (Gelfand and Smith 1990) and other Markov chain Monte Carlo techniques (see, for example, Gilks, Richardson, and Spiegelhalter 1996). However, the use of these methods in analyzing supersaturated designs may require a departure from noninformative priors to guarantee the existence of a proper posterior distribution, leading to possible controversy on the objectivity of the results.

This article details a two-stage Bayesian model selection strategy combining recent methodologies: the stochastic search variable selection method of George and McCulloch (1993, 1997) and the intrinsic Bayes factor method of Berger and Pericchi (1996a,b). This strategy keeps all possible models under consideration, provides a direct comparison between any two competing models, and provides a level of robustness akin to Bayesian analyses incorporating noninformative priors. Our major interest is to provide a reliable analysis for supersaturated designs, but, as pointed out by an associate editor, the proposed two-stage procedure can be used in analyzing any type of dataset, especially when (a) the number of independent variables is large, (b) a relatively small number of these factors are likely to be active, (c) only a relatively small number of observations can be taken, and (d) all active factors have first order effects which are at least as large as interactions and higher order effects.

This paper is organized as follows. Section 2 presents a short review of new Bayesian methodologies and, more specifically, summaries of the two methods to be combined here. Section 3 provides a step-by-step guideline for the proposed procedure and explains why this two-stage strategy is valuable and why it may offer important advantages over existing methods. The strategy is demonstrated on a well-known dataset in Section 4, and a simulation study is performed in Section 5. A synthetic dataset is analyzed in Section 6. Finally, Section 7 provides a summary and some concluding remarks.

2. BAYESIAN MODEL SELECTION METHODS

Markov chain Monte Carlo (MCMC) techniques are commonly used in Bayesian model selection procedures. George and McCulloch (1993, 1997) developed a Gibbs sampling strategy called stochastic search variable selection (SSVS) based on a mixture of normal priors. A short discussion of SSVS follows in Section 2.1. Chipman, Hamada, and

Wu (1997) employed SSVS to select significant factors in designed experiments. Chipman (1996) extended the use of SSVS to problems involving class variables, interactions, and other related predictors. Meyer and Wilkinson (1995) took an approach like SSVS but used as prior a mixture of normal and point mass at zero. Mitchell and Beauchamp (1988) presented a similar approach with the normal component replaced by a diffuse uniform prior. Carlin and Chib (1995) designed a Gibbs sampling scheme with different coefficients for each predictor across all candidate models. Raftery, Madigan, and Hoeting (1993), Hoeting, Raftery, and Madigan (1996), and Hoeting, Madigan, Raftery, and Volinsky (1998) employed a Metropolis–Hastings type sampler (see, for example, Bernardo and Smith 1994).

Frequently, Bayes factors are used to compare competing models for a dataset, although their calculation often requires the evaluation of high-dimensional integrals, which may be estimated using some efficient computational methods (Gelman and Meng 1998). One drawback, however, is that the prior distributions must be proper to ensure identifiability of the Bayes factor. Berger and Pericchi (1996a,b) developed a modification, the intrinsic Bayes factor (IBF), by using training samples to convert improper priors to proper posteriors (conditional on the training data). This allows the use of noninformative priors while maintaining identifiability. A short discussion of IBF follows in Section 2.2.

2.1 The MCMC Approach of SSVS

The SSVS setup of George and McCulloch (1993, 1997) has been widely cited and has the advantage that the random regression coefficients can be generated jointly instead of one at a time, leading to greater computational speed (George and McCulloch 1997). Although other cited strategies have their own strengths, SSVS is chosen here to analyze supersaturated designs primarily on the basis of its widespread acceptance and use in the literature.

The SSVS method starts with the usual linear model assumptions for regression,

$$(Y \mid \beta, \sigma) \sim N_n(X\beta, \sigma^2 I_n), \tag{1}$$

where X is an n by k model matrix with columns for predictors centered and scaled so that $\sum_{j=1}^{n} x_{ij}^2 = 1$, and $\beta = (\beta_1, \ldots, \beta_k)$ is a vector of coefficients for the k factors in X. The method excludes from X those terms that are included in every model, such as the intercept. To account for these terms, one replaces the Y vector with the residual vector obtained by regressing Y on those terms. Thus, if only the intercept is included in every model, one replaces Y with $Y - \overline{Y}$, where $\overline{Y}$ is the sample mean.

George and McCulloch (1993) specify a mixture prior distribution on the regression coefficients,

$$(\beta \mid \gamma) \sim N_k(0, D_\gamma R D_\gamma), \tag{2}$$

where R is the prior correlation matrix, D_γ is a diagonal matrix consisting of elements $a_1\tau_1, \ldots, a_k\tau_k$,

$$a_i = \begin{cases} 1 & \text{if } \gamma_i = 0, \\ c_i & \text{if } \gamma_i = 1, \end{cases}$$

which determines the variance of the mixture normal for values of a set of Bernoulli variable inclusion parameters $\gamma = (\gamma_1, \ldots, \gamma_k)$, and $\tau' = (\tau_1, \ldots, \tau_k), (c_1, \ldots, c_k)$ are appropriately chosen tuning constants. The choices of $\tau_1, \ldots, \tau_k, c_1, \ldots, c_k$ should be set so that the τ corresponds to a coefficient that can be safely estimated by zero, and the $c\tau$ corresponds to a typical large value that we should be able to identify. In Section 3, we will discuss specifically how to choose these variables interact with the experimental goals in the context of supersaturated designs.

We first discuss the choice of R. The correlation matrix of the regression coefficient estimates $\hat{\beta}_1, \ldots, \hat{\beta}_k$ is $(X^T X)^{-1}$, since the columns of X were appropriately centered and scaled. Such a natural choice for R does not exist in the supersaturated design setting. George and McCulloch (1993) offer an alternative choice of $R = I_k$, the identity matrix of order k, and use it in most of their examples. This choice seems to work well in practice, and the assumption that the true regression coefficients are a priori independent may be reasonable. A similar practice is adopted here.

For the prior on σ^2, George and McCulloch (1993) chose

$$\left(\frac{\nu_\gamma \lambda_\gamma}{\sigma^2} \,\middle|\, \gamma \right) \sim \chi^2_{\nu_\gamma},$$

where $\nu_\gamma \geq 0$ and $\lambda_\gamma \geq 0$ are appropriately chosen tuning constants. Finally, they specified each γ_i as an independent Bernoulli with probability π_i, so that the prior probability that the ith coefficient is practically significant is π_i. That is, $f(\gamma) = \prod_{i=1}^{k} \pi_i^{\gamma_i}(1 - \pi_i)^{1-\gamma_i}$. Here, the default "indifference" prior is $f(\gamma) = 2^{-k}$ with $\pi_i = 1/2$

The Gibbs sampling algorithm provides an approximate random sample from the posterior distribution of the inclusion parameters given the data and hence estimates of the model probabilities. We then rank all candidate models by their estimated posterior probabilities and select the one model with highest posterior probability.

Advantages of the SSVS approach include the following.

- The Gibbs sampling approach allows SSVS to keep all possible models under consideration, even though there are more parameters than observations.
- By ranking the competing models according to their frequencies in the Gibbs sample, one can narrow down the choice of top model(s) or model factors to those with the highest posterior probabilities.

On the other hand, there are a few limitations, as follows.

- Results depend on the choices of proper prior distributions, which are needed to yield a proper posterior distribution. The SSVS setup employs a mixture normal distribution for coefficients and a gamma distribution for variance components. Results of the algorithm may be highly sensitive to choices of the tuning constants that specify these distributions. The problem becomes more serious when parameters are not estimable as in the case of supersaturated designs.
- Gibbs sampling algorithms require an investigation to ensure convergence to the posterior distribution. George and McCulloch (1993) intended the use of SSVS to be exploratory

in nature, so they promoted the practice of no initializing iterations to reduce computation time. If several models have similarly large posterior probabilities, then this practice of sampling from the Markov chain before convergence has been achieved casts doubt on the correctness of the top model choice.

2.2 The Bayes Factor Approach of IBF

With k factors there are 2^k competing models—$M_1, \ldots, M_{2^k}$—for the model selection problem. Under model M_i, let the density of Y be given by $f_{Y|\Theta_i}(y \mid \theta_i)$, which depends on a (vector) parameter Θ_i. Similarly under M_j, let the density of Y be $f_{Y|\Theta_j}(y \mid \theta_j)$, which depends on Θ_j. Let $f_{\Theta_i}(\theta_i)$ and $f_{\Theta_j}(\theta_j)$ be priors for the two parameters. Using this notation, the Bayes factor for comparing model M_i to model M_j is

$$\mathrm{BF}_{ij} = \frac{\int f_{Y|\Theta_i}(y \mid \theta_i) f_{\Theta_i}(\theta_i) d\theta_i}{\int f_{Y|\Theta_j}(y \mid \theta_j) f_{\Theta_j}(\theta_j) d\theta_j}. \tag{3}$$

One way to perform a robust Bayesian analysis is to use (improper) noninformative priors for Θ_i and Θ_j, so that the likelihood function dominates the prior. However, multiplication of any improper prior by a constant yields an equivalent prior but a different value for the Bayes factor, rendering the Bayes factor useless. To correct this deficiency in standard Bayes factors, Berger and Pericchi (1996a) used training samples to convert improper prior distributions to proper posteriors (conditional on the training samples), which are then used as priors for the remaining data. The result was their IBF

$$\mathrm{IBF}_{ij}(l) = \frac{\int f_{Y(-l)|\Theta_i, Y(l)}(y(-l) \mid \theta_i, y(l)) f_{\Theta_i|Y(l)}(\theta_i \mid y(l)) d\theta_i}{\int f_{Y(-l)|\Theta_j, Y(l)}(y(-l) \mid \theta_j, y(l)) f_{\Theta_j|Y(l)}(\theta_j \mid y(l)) d\theta_j}, \tag{4}$$

where $y(l)$ is a minimal training sample and $y(-l)$ consists of the remaining data. Since the choice of minimal training sample will affect the IBF, one averages arithmetically (AIBF) or geometrically (GIBF) over all possible minimal training samples, L, to obtain a more stable value.

Advantages of the intrinsic Bayes factor approach include the following.

- The procedure directly compares the fits of any two competing models, whether nested or nonnested, and chooses the one with the highest increase in posterior odds over its prior odds.
- Although the results do depend on the parametric setup of the model (e.g., linear model with normal error) as almost all procedures do, the results are fairly robust since noninformative priors are employed.
- The procedure can be set to run automatically because no prior input is needed for the noninformative priors.

Alternatively, there are a few limitations, as follows.

- When the number of factors exceeds the number of observations as is the case in supersaturated designs, there exists no training sample that can convert an improper prior to a proper posterior. The procedure, like standard regression techniques, is computationally impossible and cannot keep all possible subsets of the predictors under consideration.
- Even if the design used yielded estimable effects (that is, the design is not supersaturated), a large pool of potential regressors renders the technique infeasible due to the large number (2^k) of comparisons to be performed.

3. A TWO-STAGE SELECTION STRATEGY

It is well known that a stepwise regression scheme tends to include far too many factors in its choice of best model (Westfall et al. 1998). A recent paper by Abraham, Chipman, and Vijayan (1999) suggests that all-subsets regression is a better alternative than stepwise regression. However, such an approach is impractical even when a moderate number of factors are active. For example, a supersaturated design with 23 factors and at most six active factors leads to possible consideration of $\sum_{i=1}^{6} \binom{23}{i} = 145,498$ models, a formidable model comparison.

The two Bayesian methods previously discussed are not suitable for the analysis of supersaturated designs, either. However, when combined into a single two-stage procedure, a powerful tool results, as will be demonstrated. One limitation of IBF is that it cannot be used on supersaturated design data, but SSVS can. One limitation of SSVS is that its sensitivity to tuning constant choices and convergence determination make it unlikely to select one best "objective" model, but the strength of IBF is that the use of a noninformative prior allows objectivity in its selection. Often one is willing to use informative priors when the number of factors is large because identifying the few good models is easier than determining the absolute best among those, for which one needs an impartial tool to remain objective. SSVS, which is an exploratory tool able to keep all possible models under consideration, can be used to make the gross comparisons and shorten the list of candidate models and factors, while IBF can be the impartial tool used to make the final model decision—a two-stage strategy designed to take full advantage of the strengths of each approach.

Some critics may argue that a more traditional approach, such as stepwise regression or all (computationally) possible subsets evaluation, be used in place of SSVS in the two-stage strategy. Although these methods might also narrow the list of candidate factors, SSVS keeps all possible models under consideration in the supersaturated design setting. An additional benefit of SSVS is its fast computational speed compared to the other "all-subsets" alternatives. A general guideline to implement the two-stage strategy is given below.

3.1 First Stage: SSVS

A. *Choices of the Tuning Constants.* Application of SSVS requires the specification of tuning constants ν_γ and λ_γ in the prior for σ^2 and c_i and τ_i ($i = 1, \ldots, k$) in the prior for β (cf. Equations (2) and (3)). George and McCulloch (1993, 1997) used $\nu_\gamma = \nu = 0$, yielding a noninformative prior for σ^2. A different choice must be made when analyzing supersaturated designs, as the use of a noninformative prior will lead to an improper posterior. The interpretation that $[\nu_\gamma/(\nu_\gamma - 2)]\lambda_\gamma$ is a fictitious prior estimate for the value of σ^2 from a sample of size ν_γ (George and McCulloch 1993) can be used to help specify these values. To obtain a diffuse proper prior, one

may select $\nu_\gamma = \nu = 3$, the smallest integer yielding a positive estimate, and $\lambda_\gamma = \lambda = \hat{\sigma}^2/3$, where $\hat{\sigma}^2$ is an estimate of σ^2, as tuning constants. Certainly, if any information about σ^2 is available and a proper prior can be specified, one can easily incorporate it into the analysis.

For (τ_i, c_i), George and McCulloch (1993) offered four semiautomatic choices: $(\hat{\sigma}_{\beta_i}, 5)$, $(\hat{\sigma}_{\beta_i}, 10)$, $(.10\hat{\sigma}_{\beta_i}, 100)$, and $(.10\hat{\sigma}_{\beta_i}, 500)$, where $\hat{\sigma}_{\beta_i}$ is the standard error of the least squares estimate of β_i. These four choices were incorporated in all simulations and example data analyses in this paper. (Later, the abbreviated notations (1, 5), (1, 10), (.10, 100), (.10, 500), respectively, will be used to refer to them.) Note that these four choices are not exhaustive; other choices of tuning constants can also be used. However, using several sets of tuning constants in the first stage to generate sufficient variables for the second stage is recommended. The attractiveness of specifying τ_i in terms of $\hat{\sigma}_{\beta_i}$ is that one can weigh practical significance, quantified by τ_i, against statistical significance, quantified by $\hat{\sigma}_{\beta_i}$, and that the choices are invariant to scaling of the predictors (George and McCulloch 1993). However, since $\hat{\sigma}_{\beta_i}$ cannot be estimated from supersaturated design data without assumptions on β, an alternative estimate will be needed and is discussed below.

B. *Starting Values.* The standard choice of starting values for the β vector and σ^2 within the Gibbs sampling algorithm is the set of full model least squares estimates—an impossible strategy for supersaturated designs. A simpler strategy is to use as estimates of β_i $(i = 1, \ldots, k)$ those coefficients obtained from fitting all k simple linear regression models. Although these estimates may be biased, the starting values are not critical to the success of the Gibbs sampler. More importantly, the estimate of β does not affect the specification of any tuning constants.

Unlike the estimate of β, an accurate estimate of σ^2 is critical to the success of the algorithm because tuning constants λ and $\tau = (\tau_1, \ldots, \tau_k)$ are defined by it. It can be shown that the conditional distribution used to generate γ_i in the Gibbs sampler is Bernoulli with probability

$$\frac{1}{1 + \frac{1-\pi_i}{\pi_i} c_i \exp\left\{-\frac{1}{2}\beta_i^2 \frac{c_i^2 - 1}{c_i^2 \tau_i^2}\right\}}$$

that factor X_i is included. Examination of this quantity confirms that large τ_i's produce models with few factors while small τ_i's produce models with many factors. Given the wide availability of statistical software for performing stepwise regression, a computationally simple alternative is to use the estimate of σ^2 from the final model selected by that method. Care should be exercised in the selection of a p-value to be used in the stepwise regression procedure, however, as stepwise techniques (using a large p-value) are prone to overfitting (leading to an underestimate of σ^2). In practice, one could make use of several different choices, but experience gained from applying this practice to the example of Section 4 and the simulations of Section 5 suggests that use of .05 for the p-value may work well.

George and McCulloch (1993, 1997) suggest that the Gibbs sampler should begin with all $\gamma_i = 1$, which corresponds to starting with the full model. This seems appropriate for supersaturated designs, too.

3.2 Second Stage: IBF

The factors identified in the SSVS stage using various choices of tuning constants are now used as the input for the IBF analysis in the second stage. Berger and Pericchi (1996a) considered priors of the form $f(\beta, \sigma) \propto \sigma^{-(1+q)}$, where q is a constant greater than -1, to be used in computing the IBF. This noninformative specification encompasses several priors for normal linear models. Although in practice one might make use of all these prior choices, Bernardo (1979) has provided a convincing argument for the use of the reference prior (with $q = 0$), which will be used within illustrative examples and simulation comparisons to follow. The procedure can be set to run automatically.

3.3 A Step-By-Step Procedure for the Proposed Method

A step-by-step guideline for the proposed procedure can be summarized as follows.

1. Identify all the candidate factors.
2. Center all the predictors and scale them so that each has sum (across observations) of squares equal to 1.
3. Run stepwise regression using a p-value criterion of $p = .05$. Using the final model that stepwise selects, obtain an estimate of the residual variance.
4. Run the SSVS procedure of George and McCulloch (1993) on the full dataset (all factors) using several different choices for tuning constants (e.g., $(c_i, \tau_i) = (\hat{\sigma}_{\beta_i}, 5)$, $(c_i, \tau_i) = (\hat{\sigma}_{\beta_i}, 10)$, $(c_i, \tau_i) = (.10\hat{\sigma}_{\beta_i}, 100)$, and $(c_i, \tau_i) = (.10\hat{\sigma}_{\beta_i}, 500)$, where $\hat{\sigma}_{\beta_i} = \hat{\sigma}$).
5. From each of these four SSVS runs, identify the model with the highest estimated posterior probability. Select only models that are distinguishable from the other sampled models. Also identify any other apparent important factors that may not have been included in the top model, such as factors that exist in 5 of the top 10 sampled models.
6. Combine all the models and factors selected in the previous step into one "encompassing" model.
7. Run the IBF procedure of Berger and Pericchi (1996a) on the encompassing model using the reference prior (or with several different prior choices).
8. Identify the best model selected by IBF using either AIBF or GIBF averaging.
9. Proceed with regression diagnostics on the final model choice to assure oneself of the accuracy of the final selection.

4. A NEW LOOK AT WILLIAMS' DATA

Lin (1993) presented an analysis of a half-fraction Plackett–Burman type design originally published in Williams (1968). The design investigated 24 factors in 14 observations, although 2 of the factors in the original dataset, 13 and 16, were completely confounded. Table 1 reproduces this dataset, with factor 13 deleted and factors 14–24 renamed as 13–23, respectively.

Using his entire dataset (two half-fractions), Williams (1968) identified factors 4, 10, 14, and 19 as the most influential. Lin (1993) utilized a forward selection procedure on this half-fraction for identifying the most important factors.

Table 1. Williams' Half-Fraction Published in Lin

	Run													
Factor	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*
1	+	+	+	+	−	−	−	−	−	+	−	+	+	−
2	+	−	+	+	−	−	−	+	−	+	+	−	+	−
3	+	−	−	−	+	+	−	+	−	+	−	−	+	+
4	−	−	+	+	+	+	−	−	−	+	+	−	+	−
5	−	−	+	−	+	+	+	−	−	−	+	+	+	−
6	−	−	−	+	+	+	−	+	+	+	−	+	−	−
7	+	+	−	−	−	+	−	−	+	+	−	+	+	−
8	+	+	−	−	+	−	+	+	−	+	+	−	−	−
9	+	+	−	−	+	+	−	−	−	−	+	+	+	−
10	+	−	+	+	−	+	+	+	−	−	−	+	−	−
11	+	−	−	+	−	+	−	−	+	−	+	+	−	+
12	−	−	+	−	−	−	+	−	+	+	−	+	+	+
13	−	+	+	−	−	+	+	−	−	+	+	−	−	+
14	−	+	+	+	+	+	−	−	+	+	−	−	−	−
15	+	+	+	+	+	−	+	−	−	−	−	+	−	−
16	+	−	+	−	+	+	+	−	+	+	−	−	−	−
17	−	+	−	+	−	+	+	−	+	−	+	−	+	−
18	−	−	−	+	−	+	+	+	−	+	+	+	−	−
19	+	−	−	+	+	+	+	−	−	−	−	−	+	+
20	−	+	−	−	−	+	+	+	−	+	−	+	+	−
21	−	+	+	−	+	+	−	+	−	−	−	+	−	+
22	−	−	+	−	+	−	−	+	+	−	+	+	+	−
23	+	−	−	−	+	−	+	−	+	+	+	+	−	−
y	133	62	45	52	56	47	88	193	32	53	276	145	130	127

His analysis also identified these 4 factors in addition to factor 12. Noting that sequential selection procedures are prone to overfitting, Westfall et al. (1998) used resampling to estimate the distribution of the maximal F statistic at each step of the procedure. They used those estimates to adjust the p-value cutoffs for inclusion in forward selection. Using their methodology on the Williams half-fraction dataset, they concluded that only factor 14 was active.

Several other stepwise analyses were carried out on the half-fraction. Using a cutoff p-value of .05 (or .01), both forward and stepwise regression (implemented in SAS, SAS Institute Inc. 1994) selected only factor 14. Increasing the criterion to .10 caused 12 factors—the same for forward as for stepwise—to be flagged as important. A stepwise procedure utilizing the AIC criterion (implemented in S-Plus, Statistical Sciences 1995) selected factor 14 only.

The SSVS algorithm was also run, using the tuning constants and starting values as outlined in Section 3. The implementation of SSVS used the σ^2 estimate of $(43.88)^2$ obtained from stepwise regression with $p = .05$. Some of the top identified models for each choice of (τ_i, c_i) in Section 3.1 are presented with their model probabilities in Table 2. The model probabilities are all near zero and nearly equal when using $(1, 5)$; it is doubtful that this result sheds any light on the best model, although its top choice—factor 14 only—is the same as that selected by (1, 10) and (.10, 500). In

Table 2. Top Models Selected by SSVS Using Stepwise Regression (p-Value = .05) Error Variance Estimate

$\tau_i = \hat{\sigma}_{\beta_i}$, $c_i = 5$		$\tau_i = \hat{\sigma}_{\beta_i}$, $c_i = 10$	
Prob.	*Model*	*Prob.*	*Model*
.0010	14	.0170	14
.0008	13, 14, 16, 19	.0058	14, 19
.0008	8, 11, 14, 22	.0048	12, 14
.0008	11, 14, 19	.0044	2, 14
.0006	(eight models)	.0036	(three models)
$\tau_i = .10\hat{\sigma}_{\beta_i}$, $c_i = 100$		$\tau_i = .10\hat{\sigma}_{\beta_i}$, $c_i = 500$	
Prob.	*Model*	*Prob.*	*Model*
.1142	4, 12, 14, 19	.2288	14
.0268	4, 10, 12, 14, 19	.0792	4, 12, 14, 19
.0188	4, 12, 14, 19, 20	.0766	12, 14, 19
.0186	1, 4, 12, 14, 19	.0618	12, 14
.0150	4, 11, 12, 14, 19	.0268	14, 16

contrast, the model with factors 4, 12, 14, and 19 was chosen best under (.10, 100). Strong runners-up included several submodels of {4, 12, 14, 19}, namely itself, {12, 14, 19} and {12, 14} when using (.10, 500), and {14, 19} when using (1, 10).

Based on the selections of the SSVS algorithm, {4, 12, 14, 19} was chosen as the model for input into the IBF scheme. Table 3 lists the intrinsic Bayes factors for comparing this encompassing model to each submodel, arranged in order of increasing AIBF. The best model selected by this approach was the encompassing model. Second on the list was the model consisting of {12, 14, 19}. This submodel had an AIBF of 31.4 (GIBF of 4.47), meaning that the odds in favor of the encompassing model (compared to {12, 14, 19}) have increased 31.4 (4.47) times by accounting for the data. This is strong evidence in favor of {4, 12, 14, 19}. Beattie (1999) has implemented SSVS using an error variance estimate from stepwise regression with $p = .10$, yielding top models involving more than 10 factors. The IBF follow-up procedure, however, selected {14, 22}. A further analysis using {4, 12, 14, 19, 22} as the encompassing model concludes that {4, 12, 14, 19} is the best model (see Beattie 1999 for details).

Table 4 summarizes the results of the various analytical techniques applied to the Williams half-fraction. The classical stepwise techniques tend to be either very conservative (using $p = .01$ or .05), selecting one factor, or very liberal (with $p = .10$), selecting 12 factors (not shown in the table). SSVS, when using one estimate of σ^2, agrees closely with the conservative stepwise methods but while using a different estimate is extremely liberal in factor inclusion. The IBF follow-up procedure to SSVS results in a more moderate, stable final model, with four factors selected from the factors identified by SSVS. Note that the result from the two-stage procedure is *insensitive* to the estimate of σ^2 and is close to the final model suggested by Williams (1968). Using the all-subsets regression procedure recommended by Abraham, Chipman, and Vijayan (1999), assuming five or fewer active factors, yields the same four-factor model as the two-stage procedure. However, our Bayesian approach does not need to specify the number of active factors in advance.

Table 3. Intrinsic Bayes Factors of {4,12,14,19} Against All Its Subsets

Model factors	AIBF	GIBF
4, 12, 14, 19	1.00	1.00
12, 14, 19	31.4	4.47
4, 14, 19	94.7	23.2
4, 12, 14	177	46.2
12, 14	378	18.8
14, 19	693	25.1
4, 14	1050	69.1
14	3420	30.7
None	4880	322
12	6020	611
19	8220	666
4	8900	839
12, 19	10500	1550
4, 12	11600	1700
4, 19	15300	2080
4, 12, 19	24200	6190

Table 4. Comparative Results for Analyses of the Williams Half-Fraction

Model Selection Method	Factors Identified as Important
Williams' final model (Williams 1968)	4, 10, 14, 19
Forward selection w/modified p (Westfall et al. 1998)	14
Forward selection w/$p = .05$	14
Stepwise w/$p = .05$	14
Stepwise using AIC	14
SSVS ($\tau_i = \hat{\sigma}_{\beta_i}$, $c_i = 5$)	14
SSVS ($\tau_i = \hat{\sigma}_{\beta_i}$, $c_i = 10$)	14
SSVS ($\tau_i = .10\hat{\sigma}_{\beta_i}$, $c_i = 100$)	4, 12, 14, 19
SSVS ($\tau_i = .10\hat{\sigma}_{\beta_i}$, $c_i = 500$)	14
IBF using 4, 12, 14, 19 from SSVS	4, 12, 14, 19

5. SOME COMPARISONS BY SIMULATION

In practice, the two-stage approach would be carried out in the manner of Section 4, with careful selection of the encompassing model for IBF occuring by examination of a few top model or factor choices from the four SSVS runs using different estimates of σ^2. For simulation, such an in-depth process is not feasible. For purposes of this study, the best model with most factors selected by a specific SSVS run was used as the encompassing model for IBF, which compared all subsets of this SSVS choice. For cases in which the SSVS top selection contained too many factors for IBF, the top six marginal factor probabilities were used to obtain the encompassing model. Note that if the SSVS top selection contains too many factors, this could also be evidence that the assumption of effect sparsity is unrealistic. Indeed, when the effect sparsity assumption is questionable, the supersaturated design may not be appropriate and more observations (experiments) are needed. Because of these limitations, it is believed that the simulation studies will understate the success of the two-stage strategy. All priors for IBF, as suggested in Berger and Pericchi (1996a), were used here, although only the results under the reference prior will be presented. The results for the other priors did not differ greatly from those of the reference prior.

In the model selection setting, there are two different goals for the selection algorithm. Ideally, the technique should correctly identify all the true active factors and ignore those that are inactive. On the other hand, supersaturated designs are often used as first-line screening procedures, in which one seeks to correctly identify all true active factors but may tolerate the inclusion of a few inactive ones. The number of inactive factors included should be kept to a minimum so that follow-up experiments do not get overly expensive or unwieldy. With these two goals in mind, the procedures were judged by two criteria: (1) how well they could identify the true model and (2) how well they could identify the true active factors yet keep fitted model size to a minimum.

Simulation studies on several randomly generated datasets were used to evaluate the two-stage Bayesian model selection strategy against stepwise regression and SSVS. The first simulation investigated the abilities of the selection techniques in the presence of one active factor, namely $10X_1$,

Table 5. Percent of 1000 Simulations Succeeding in Identi' cation

True Model: $Y \sim N(10X_1, I_{14})$

Model Selection Method	True Model Identi' ed	Active Factor Identi' ed	Average Size	
			Med.	Mean
Stepwise ($p = .05$)	22.7%	100%	3	3.2
Stepwise (AIC)	99.9%	100%	1	1.0
SSVS (.10, 500)	40.5%	99%	2	3.1
Stepwise/GIBF	24.0%	100%	2	3.0
SSVS (.10, 500)/GIBF	61.0%	98%	1	2.5

True model: $Y \sim N(-15X_1 + 12X_5 - 8X_9 + 6X_{13} - 2X_{17}, I_{14})$

Model Selection Method	True Model Identi' ed	Active Factors Identi' ed	Average Size	
			Med.	Mean
Stepwise ($p = .05$)	32.9%	100%	6	6.5
Stepwise (AIC)	0%	0%–100%	1	1.2
SSVS (.10, 500)	36.4%	84%–98%	6	8.0
Stepwise/GIBF	38.3%	99%–100%	6	6.3
SSVS (.10, 500)/GIBF	40.7%	75%–94%	5	5.6

True model: $Y \sim N(-15X_1 + 8X_5 - 6X_9 + 3X_5X_9, I_{14})$

Model Selection Method	All Main Effects Identi' ed	Active Main Effects Identi' ed	Average Size	
			Med.	Mean
Stepwise ($p = .05$)	24.5%	99%–100%	5	4.9
Stepwise (AIC)	0%	0%–100%	1	1.0
SSVS (.10, 500)	38.8%	83%–98%	5	5.4
Stepwise/GIBF	29.5%	98%–100%	5	4.6
SSVS (.10, 500)/GIBF	46.5%	81%–97%	4	4.9

to gauge the amount of improvement that is possible in using the two-stage procedure over the other methods. The second simulation studied the case with five active factors $(-15X_1 + 12X_5 - 8X_9 + 6X_{13} - 2X_{17})$ so that the methods could be compared on a more typical dataset—with strong signal, moderate signal, and weak signal. The third simulation incorporated a model with three active main-effect factors and one active interaction effect $(-15X_1 + 8X_5 - 6X_9 + 3X_5X_9)$. For all simulations, the Williams half-fraction supersaturated design of Table 1 was used with responses randomly generated from a normal distribution with specified mean and variance one. Each simulation was run 1,000 times.

Table 5 displays a representative portion of the simulation results. For each model selection scheme, the percentages of correct model and factor identification are listed, along with the average model size. Stepwise results are shown for $p = .05$ and for AIC, as these are typical of selections used in practice. The two-stage procedure is performed in two ways: using stepwise as the first stage (stepwise/IBF) and using SSVS as the first stage (SSVS/IBF). Note that stepwise/IBF uniformly dominates stepwise in every criterion for almost all situations. In general, the two-stage procedure SSVS/IBF demonstrated a higher percentage of identification of the true model and a smaller (and closer to true) model size. All methods except stepwise (AIC) have similar percentage of including all active factors in the final model.

As expected among the classical stepwise procedures, an increase in the ability to identify the active factors accompanied an increase in model size (hence a decrease in identification of the correct model). The stepwise procedures generally tended to overfit the model with $p = .05$ and .10 and underfit the model with $p = .01$, although there was some model dependence on overfitting. The stepwise scheme employing the AIC criterion performed best when there was only one active factor but performed miserably when there were more than one active factor. Although stepwise/IBF dominates stepwise, the improvement is only marginal because the second stage usually confirmed all the factors identified by the stepwise procedure. On the other hand, the SSVS/IBF procedure might provide improvements because more factors may be identified in the SSVS step. Even under the severe restriction described at the beginning of this section, SSVS (.10, 500)/GIBF is able to identify the true model or all main effects at a much higher percentage as compared to the stepwise procedure.

We note that sensitivity of the tuning constants in the first stage SSVS is unlikely to affect the effectiveness of the two-stage procedure. This is because one will use all variables identified in the first stage, using various sets of tuning constants, as inputs for the second stage. Results from SSVS (first stage) depended dramatically on the choice of tuning constants but the choice of the tuning constants may not have much impact on the two-stage procedure—it will only affect the size of input variables for the second stage.

6. A SYNTHETIC DATA EXAMPLE

To demonstrate the performance of the two-stage procedure in general, a synthetic dataset generated from another type of design (not a supersaturated design) is considered here. This type of example was given in George and McCulloch (1993). We construct $k = 20$ predictor

Table 6. Comparative Results for Analyses of the Synthetic Dataset

Model selection method	*Factors identi' ed as important*
Correct model	6–20
Stepwise w / $p = .01$	11, 13, 16–20
Stepwise w / $p = .05$	5–20
Stepwise using AIC	16, 17, 19, 20
SSVS ($\tau_i = \hat{\sigma}_{\beta_i}$, $c_i = 5$)	7–20
SSVS ($\tau_i = \hat{\sigma}_{\beta_i}$, $c_i = 10$)	7–20
SSVS ($\tau_i = .10\hat{\sigma}_{\beta_i}$, $c_i = 100$)	7–20
SSVS ($\tau_i = .10\hat{\sigma}_{\beta_i}$, $c_i = 500$)	7, 9, 11–20
SSVS/AIBF	5–20
SSVS/GIBF	7–20

variables, $X_1, X_2, \ldots, X_{20}$, of length $n = 40$, obtained by $X_i = X_i^* + Z$, where X_i^*'s $\sim$ iid $N_{40}(0, I_{40})$ for $i = 1, \ldots, 20$, and are independent of $Z \sim N_{40}(0, I_{40})$. The response variable was generated by the model $\mathbf{Y} = [X_1, \ldots, X_{20}]\beta + \epsilon$, where $\epsilon \sim N_{40}(0, \sigma^2 I_{40})$ with $\sigma = 2$ and the coefficients $\beta' = (\beta_1, \ldots, \beta_{20})$ were set at $(\beta_1, \ldots, \beta_5) = (0, 0, 0, 0, 0)$, $(\beta_6, \ldots, \beta_{10}) = (1, 1, 1, 1, 1)$, $(\beta_{11}, \ldots, \beta_{15}) = (2, 2, 2, 2, 2)$, and $(\beta_{16}, \ldots, \beta_{20}) = (3, 3, 3, 3, 3)$. We analyzed the generated dataset using the various methods discussed above. Results from stepwise regression employing different p-values are given in Table 6. Consistent with our simulation study results reported in Section 5, the stepwise procedure (slightly) overfitted the model with $p = .05$ and underfitted the model with $p = .01$; stepwise (AIC) did not perform well. For a clear comparison, the results from SSVS and the two-stage procedure are also included in Table 6.

Based on the selections of the SSVS algorithm, Variables 5–20 were chosen to form the encompassing model for input into the IBF scheme. Using the AIBF, the best model selected includes Variables 5–20, whereas the best model employing GIBF includes Variables 7–20. In this case, our two-stage procedure suggested the need for more observations to ascertain the significance of Variables 5 and 6. The overall conclusion, however, is satisfactorily accurate.

7. SUMMARY AND CONCLUSION

Experiments can be costly and involve a large number of factors. When only a few factors of which are thought to be active, a supersaturated design can be gainfully employed. However, the use of supersaturated designs for screening experiments in industry has brought about the need for new model selection methodologies. Classical linear regression procedures are unable to estimate all the parameters. Stepwise techniques, although appealing for their automation, are prone to overfitting. More significantly, the selection of a p-value cutoff may be tied to the true, but unknown, underlying model. On the other hand, the use of alternative stepwise criteria, such as AIC or C_p, can lead to underfitting when more than one factor is active.

Abraham, Chipman, and Vijayan (1999) also pointed out some shortcomings of stepwise regression for analyzing supersaturated designs and favored the use of all-subsets regression. Some of their observations were in fact given in Lin (1995b). Their preferred all-subsets regression procedure, however, is viable for models with only a few active factors. Williams' example involves 23 candidate factors and $\sum_{i=1}^{5} \binom{23}{i} = 44,551$ models for five or fewer active factors, and 145,498 models when six or fewer active factors are to be considered. For a design with even a moderate number of active factors under study, such a procedure would be formidable (from the computational perspective) to implement.

Bayesian MCMC techniques are attractive because they can keep all 2^k models under consideration and are computationally feasible. The SSVS procedure of George and McCulloch (1993, 1997) is one popular tool for model selection. However, its dependence on proper prior distributions results in estimated model probabilities that depend greatly on the selection of tuning constants. It is unlikely that only one set of tuning constants would be appropriate for every underlying model. In practice, one would consider the results from a variety of tuning constant values; yet this approach would not necessarily lead to a single choice of best model. This strategy would also seem to suffer from overfitting as the candidate models and factors are combined into one encompassing model. Instead of finding the one best model, the methodology at best identifies a set of candidate models or factors for further consideration.

The intrinsic Bayes factor of Berger and Pericchi (1996a,b) is appealing on theoretical considerations and because of its use of noninformative prior distributions. However, when used by itself in analyzing supersaturated design data, it suffers from nonidentifiability of the parameter values. When the number of candidate factors is reduced to a reasonable amount, as in the proposed two-stage Bayesian model selection strategy, IBF shows an ability to further reduce the model to a moderate size and to do so accurately. Our major interest here is to provide a reliable analysis for supersaturated designs, but the proposed two-stage procedure can also be used in analyzing any type of dataset, when (a) the number of independent variables is large, (b) a relatively small number of these factors are likely to be active, (c) only a relatively small number of observations can be taken, and (d) all active factors have first order effects which are at least as large as interactions and higher order effects.

ACKNOWLEDGMENTS

We thank the editor, an associate editor, and two referees for their thorough and constructive comments. We also thank Dr. Robert McCulloch for providing us with a C^{++} program to perform his SSVS methodology and Dr. Luis Pericchi for giving us a copy of his Fortran program to implement IBF. Their programs were used to independently verify the accuracy of our C program for the two-stage Bayesian model selection. Dennis Lin is partially supported by the U.S. National Science Foundation via grant DMS-9704711 and the National Science Council of ROC via contract NSC 90-2118-M-001-010.

[Received January 1999. Revised March 2001.]

REFERENCES

Abraham, B., Chipman, H., and Vijayan, K. (1999), "Some Risks in the Construction and Analysis of Supersaturated Designs," *Technometrics*, 41, 135–141.

Beattie, S. D. (1999), "Contributions to the Design and Analysis of Experiments," Ph.D. thesis, The Pennsylvania State University.

Berger, J., and Sellke, T. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of p-Values and Evidence," *Journal of the American Statistical Association*, 82, 112–122.

Berger, J. O., and Pericchi, L. R. (1996a), "The Intrinsic Bayes Factor for Linear Models," in *Bayesian Statistics—Proceedings of the 5th Valencia International Meeting Held in Alicante, June 5–9, 1994*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford Science Publications, Vol 5, pp. 25–44.

———. (1996b), "The Intrinsic Bayes Factor for Model Selection and Prediction," *Journal of the American Statistical Association*, 91, 109–122.

Bernardo, J.-M. (1979), "Reference Posterior Distribution for Bayesian Inference," *Journal of the Royal Statistical Society*, Ser. B, 41, 113–147.

Bernardo, J.-M., and Smith, A. F. M. (1994), *Bayesian Theory*, New York: Wiley.

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

Carlin, B. P., and Chib, S. (1995), "Bayesian Model Choice via Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society*, Ser. B, 57, 473–484.

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

Chipman, H., Hamada, M., and Wu, C. F. J. (1997), "A Bayesian Variable Selection Approach for Analyzing Designed Experiments With Complex Aliasing," *Technometrics*, 39, 372–381.

Draper, N. R., and Smith, H. (1998), *Applied Regression Analysis* (3rd ed.), New York: Wiley.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A., and Meng, X. (1998), "Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling," *Statistical Science*, 13, 163–185.

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

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

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (eds.) (1996), *Markov Chain Monte Carlo in Practice*, London: Chapman & Hall.

Hoeting, J., Raftery, A. E., and Madigan, D. (1996), "A Method for Simultaneous Variable Selection and Outlier Identification in Linear Regression," *Computational Statistics and Data Analysis*, 22, 251–270.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1998), "Bayesian Model Averaging," Technical Report 335, University of Washington, Department of Statistics.

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

———. (1995a), "Generating Systematic Supersaturated Designs," *Technometrics*, 37, 213–225.

———. (1995b), Response to "Letter to the Editor," by Wang, *Technometrics*, 37, 359.

———. (1999), "Supersaturated Designs," in *Encyclopedia of Statistical Sciences*, updated Volume 3, New York: Wiley, pp. 727–731.

Meyer, R. D., and Wilkinson, R. G. (1995), "Variable Selection or Variable Assessment?" Technical Report 126, University of Wisconsin–Madison, Center for Quality and Productivity Improvement.

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

Pan, G. (1999), "The Impact of Unidentified Location Effects on Dispersion-Effects Identification from Unreplicated Factorial Designs," *Technometrics*, 41, 313–326.

Raftery, A., Madigan, D., and Hoeting, J. (1993), "Model Selection and Accounting for Model Uncertainty in Linear Regression Models," Technical Report 262, University of Washington, Department of Statistics, to appear in *Journal of the American Statistical Association*, 92(437), 179–191.

SAS Institute Inc. (1994), *SAS/STAT User's Guide, Version 6* (4th ed.), Cary, NC: SAS Institute Inc.

Statistical Sciences (1995), *S-PLUS Guide to Statistical and Mathematical Analysis* (3.3 ed.), Seattle: StatSci, a division of MathSoft, Inc.

Westfall, P. H., Young, S. S., and Lin, D. K. J. (1998), "Forward Selection Error Control in the Analysis of Supersaturated Designs," "*Statistica Sinica*, 8, 101–117.

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