# Bayesian prior distributions for the analysis of screening experiments

Hugh Chipman (University of Waterloo, Canada)

May 6, 2004

**Abstract**

The use of screening experiments to discover active effects often leads to some compromises. When many effects are under consideration, designs may have complex aliasing patterns between effects. This is especially the case when interactions and higher-order effects exist. Complex aliasing and a large number of candidate effects makes selection of subsets a challenging problem. This chapter outlines Bayesian methods for subset selection, with an emphasis on the choice of prior distributions and their impact on subset selection, computation and practical analysis. Attention focuses on the linear regression model with Gaussian errors, although extensions are discussed. A number of advantages of the Bayesian approach are stressed, such as the ability to incorporate useful information about subset preferences. For example, an $AB$ interaction effect might only be considered active if main effects for $A$ and $B$ are also likely to be active. When such information is combined with a stochastic search for promising subsets, a powerful subset selection tool results. These techniques may also be applied to designs without complex aliasing as a way of quantifying subset uncertainty.

**Keywords:** Effect heredity, linear regression, subset selection, model selection, interaction effect, Markov chain Monte Carlo.

## 1 Introduction

Many of the ideas in this chapter are best motivated by example. The introduction will start with a motivating example, followed by an overview of the rest of the chapter.

Table 1: 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 |
| $D$ | Speed of centrifuge | 2100, 2300, 2500 RPM |
| $E$ | Time in centrifuge | 1.75, 3.00, 4.50 minutes |
| $F$ | (Sensitivity, Absorption) | (0.10,2.5), (0.25,2.0), (0.50,1.5) |
| $G$ | Temperature | 25, 30, 37 $^{\circ}C$ |
| $H$ | Dilution | 1:51, 1:101, 1:151 |

### 1.1 Motivating example: a blood-glucose experiment with mixed-level design

Henkin (1986) used an 18-run mixed-level design to study the effect of 8 factors on blood-glucose readings made by a clinical laboratory testing device. Factor names and settings are given in Table 1, and the design is given in Table 2. One factor, $A$ is set at two levels, while each of the other 7 $(B - H)$ are set at three levels.

A goal of screening designs such as this glucose experiment is to identify the active effects. The general approach outlined in this chapter is to view the analysis of screening data as a regression problem, in which an $n \times p$ matrix $X$ of predictors is constructed using the levels of the factors. For example, in the glucose data, linear and quadratic main effects and interaction effects between them will be considered. These effects will be represented by a vector of regression coefficients $\boldsymbol{\beta}$, and the corresponding columns of $X$ will be referred to as contrasts. Most of the effects will be assumed to be near zero (*inactive*). The task of identifying a subset of active effects corresponds to identifying which contrasts in $X$ should be included in a regression model. Box & Meyer (1986) introduced the concept of the activity of an effect, and presented one of the first Bayesian methods for the analysis of designed experiments.

The selection of individual effects may seem counterintuitive for screening experiments, since the primary goal is to identify important factors. The philosophy behind identification of individual active effects is that the once active effects are identified, the corresponding factors will also be known. An emphasis on selection of active effects instead of active factors means that more insight is gained into the nature of the relationship between the factors and the response.

Before illustrating procedures for the selection of subsets of active effects in the glucose data,

Table 2: Blood glucose experiment with mixed-level design and response data

| design | | | | | | | | mean |
|---|---|---|---|---|---|---|---|---|
| A | G | B | C | D | E | F | H | reading |
| 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 | 3 | 1 | 1 | 2 | 2 | 3 | 3 | 88.86 |
| 1 | 3 | 2 | 2 | 3 | 3 | 1 | 1 | 106.58 |
| 1 | 3 | 3 | 3 | 1 | 1 | 2 | 2 | 89.57 |
| 1 | 2 | 1 | 2 | 1 | 3 | 2 | 3 | 91.98 |
| 1 | 2 | 2 | 3 | 2 | 1 | 3 | 1 | 98.41 |
| 1 | 2 | 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 | 3 | 1 | 2 | 3 | 1 | 3 | 2 | 115.52 |
| 2 | 3 | 2 | 3 | 1 | 2 | 1 | 3 | 94.89 |
| 2 | 3 | 3 | 1 | 2 | 3 | 2 | 1 | 94.70 |
| 2 | 2 | 1 | 3 | 2 | 3 | 1 | 2 | 121.62 |
| 2 | 2 | 2 | 1 | 3 | 1 | 2 | 3 | 93.86 |
| 2 | 2 | 3 | 2 | 1 | 2 | 3 | 1 | 96.10 |

some details on the coding of contrasts is provided. All the factors in the glucose experiment are quantitative, making it possible to decompose the main effects for 3-level factors into linear and quadratic effects. Factor $A$ is set at 2 levels, so its linear contrast is coded as $\pm 1$. For three-level factors $(B, D, F, H)$ set at evenly spaced intervals, orthogonal polynomials are constructed. Draper & Smith (1998) provide an overview of orthogonal polynomials. Linear contrasts are coded as $(-1, 0, 1)/\sqrt{2}$ and quadratic contrasts as $(1, -2, 1)/\sqrt{6}$. Factor $F$ combines two variables (sensitivity and absorption), but since the levels of both variables are evenly spaced, $F$ will be treated as a single evenly spaced factor. For unevenly spaced factors $(C, E, G)$, contrasts are constructed differently. Linear contrasts are coded as the original values minus their means. Quadratic contrasts are formed by squaring the centered linear contrasts and then centering the squared term. So for factor $C$ with levels $\{20, 28, 35\}$, the linear contrast $C_L$ is coded as (-7.67, 0.33, 7.33), and the quadratic contrast $C_Q$ is coded as (21.22, -37.44, 16.22).

The nonregular nature of the design makes it possible to also consider interaction effects. An interaction between 2 main effects, each with 3 levels, would have 4 degrees of freedom. These will be represented by linear×linear , linear×quadratic and quadratic×quadratic effects. The corresponding contrasts are formed by multiplying two contrasts.

The contrasts just described are not scaled to be directly comparable. Since active effects are identified via regression modeling, detection of activity will be unaffected by scaling of contrasts.

A total of 113 effects are under consideration. This includes 8 linear effects $(A_L, \ldots, H_L)$, 7 quadratic effects $(B_Q, \ldots, H_Q)$, $\binom{8}{2} = 28$ linear×linear interactions $(A_L B_L, \ldots, G_L H_L)$, $7 + 7 \times 6 = 49$ linear×quadratic interactions $(A_L B_Q, \ldots, A_L H_Q, B_L C_Q, \ldots, G_L H_Q)$ and $\binom{7}{2} = 21$ quadratic×quadratic interactions $(B_Q C_Q, \ldots, G_Q H_Q)$.

With only 18 runs in the experiment, it will be important to identify a small subset of active effects from the 113 under consideration. By taking a regression approach to the analysis of this screening data, the problem reduces to one of subset selection. Hamada & Wu (1992) tackle the subset selection problem for screening experiments with a modified stepwise regression procedure. They first identify active main effects, and then identify active interactions between those active main effects. In the glucose experiment, the subset of active effects they identified was $E_Q, F_Q$ and an $E_L F_L$ interaction. This model has an $R^2$ of 68%.

Although stepwise selection algorithms run quickly, they do not consider all possible subsets. Instead, they build up a model one term at a time. An additional problem with the stepwise approach of Hamada & Wu (1992) is that by dividing the algorithm into stages, the search is further restricted: a highly significant interaction with no corresponding main effects that are active will not be identified.

A remedy to the limited scope of a stepwise search is all subsets regression, in which regression models using every possible subset of active effects are considered. Furnival & Wilson (1974)

Table 3: Glucose example: best-fitting subsets of size 1-6, identified by all-subsets search

| Subset of active effects | $R^2$ |
|---|---|
| $B_L H_Q$ | 46.2 |
| $B_L H_Q, B_Q H_Q$ | 77.0 |
| $B_L, B_L H_Q, B_Q H_Q$ | 85.5 |
| $E_Q, A_L C_L, B_L H_Q, B_Q H_Q$ | 94.3 |
| $F_L, A_L H_Q, G_L H_Q, B_L H_Q, E_L F_L$ | 97.0 |
| $A_L, E_Q, A_L C_L, B_L D_L, B_L H_Q, B_Q H_Q$ | 98.7 |

developed an efficient algorithm for this computation. For the glucose data, a search over all subsets with six or less active effects was carried out, using the `leaps` package in R (R Development Core Team 2004). The computation took about 50 minutes on a 1GHz Pentium-III. The best subset of each size is shown in Table 3. The three-term model with effects $B_L, B_L H_Q$ and $B_Q H_Q$ has $R^2 = 85.5\%$ compared to $R^2 = 68\%$ for the Hamada-Wu model.

A problem with the all subsets approach is that all relationships between predictors are ignored. For example the best model of size 4, $(E_Q, A_L C_L, B_L H_Q, B_Q H_Q)$, contains an interaction involving factors $A$ and $C$, but no corresponding main effects. Indeed, one of the main strengths of the Hamada-Wu approach is the incorporation of the principle of *effect heredity*: An interaction between two effects should not be considered active unless at least one of the corresponding main effects is also active.

Bayesian priors, coupled with efficient stochastic search algorithms, provide one approach that solves both the problem of a limited search (such as a stepwise method) and the need for constraints such as effect heredity (which are ignored by all subsets searches). To give a flavor of the Bayesian approach, two summaries of the Bayesian analysis of the glucose data are presented in this section. The analysis of the glucose data, including prior distributions used and computational methods, is described in Section 5.2.

Table 4 lists the 10 most probable subsets found by the Bayesian procedure. With the exception of the second subset listed $(B_L H_Q, B_Q H_Q)$, every term in every subset has at least one lower-order effect also in the subset. For example, in the fifth model, ($F_L$, $H_L$, $H_Q$, $A_L H_Q$, $G_L H_Q$, $B_L H_Q$, $E_L F_L$), the active effect $G_L H_Q$ has "parent" $H_Q$, which in turn has parent $H_L$. The exact notion of parents and effect heredity is stated precisely in Section 2.2.2. The fifth subset in Table 4 contains all the effects in the best subset of size 5, listed in Table 3. The Bayesian procedure has found a subset similar to one of the the best subsets, but which obeys effect heredity.

Table 4: *Glucose data: Ten subsets with largest posterior probability.*

| Subset | Posterior<br>Probability | $R^2$ |
|---|---|---|
| $B_L, B_LH_L, B_LH_Q, B_QH_Q$ | 0.126 | 86.0 |
| $B_LH_Q, B_QH_Q$ | 0.069 | 77.0 |
| $B_L, B_LH_Q, B_QH_Q$ | 0.064 | 85.5 |
| $B_L, B_LH_L, B_QH_L, B_LH_Q, B_QH_Q$ | 0.037 | 88.3 |
| $F_L, H_L, H_Q, A_LH_Q, G_LH_Q, B_LH_Q, E_LF_L$ | 0.037 | 97.0 |
| $A_L, B_L, A_LD_Q, B_LH_L, B_LH_Q, B_QH_Q$ | 0.020 | 95.0 |
| $F_L, H_L, H_Q, A_LH_Q, B_LH_Q, B_QH_Q$ | 0.019 | 93.0 |
| $F_L, H_L, H_Q, A_LH_Q, B_LH_Q, B_QH_Q, C_LH_Q$ | 0.019 | 95.9 |
| $B_L, B_Q, B_LH_L, B_LH_Q, B_QH_Q$ | 0.018 | 89.2 |
| $F_L, H_L, H_Q, A_LH_Q, G_LH_Q, B_LH_Q, B_QH_Q, E_LF_L$ | 0.017 | 97.8 |

The Bayesian approach is more than a tool to fix all subsets results by adding appropriate effects. Take for example the sixth model in Table 4: $(A_L, B_L, A_LD_Q, B_LH_L, B_LH_Q, B_QH_Q)$. The $A_LD_Q$ effect identified as part of this model does not appear in the best subsets of size 1-6. The Bayesian procedure has discovered another possible subset of effects that describes the data.

The second summary, Figure 1, displays the marginal posterior probability that each of the 113 effects are active. Effects are ordered by this probability along the horizontal axis. The vertical line for each effect represents the posterior probability of activity under a certain choice of prior hyperparameters. Robustness of these probabilities is indicated by the rectangles, which give minimum and maximum marginal posterior probability of activity over 18 different choices of prior hyperparameters. From this plot, it evident that effects $B_LH_Q$ and $B_QH_Q$ are very likely to be active, as well as other effects involving $B_L, H_L$ and $H_Q$. In addition, effects $A_LH_Q$ and $F_L$ might be active.

## 1.2 Overview of the chapter

The Bayesian approach just described can be applied to a wide variety of screening problems, including those with both quantitative and qualitative factors. Although a large number of high-order polynomials were considered in the glucose example, this technique will work equally well with linear main effects and linear×linear interactions. When there are complex aliasing

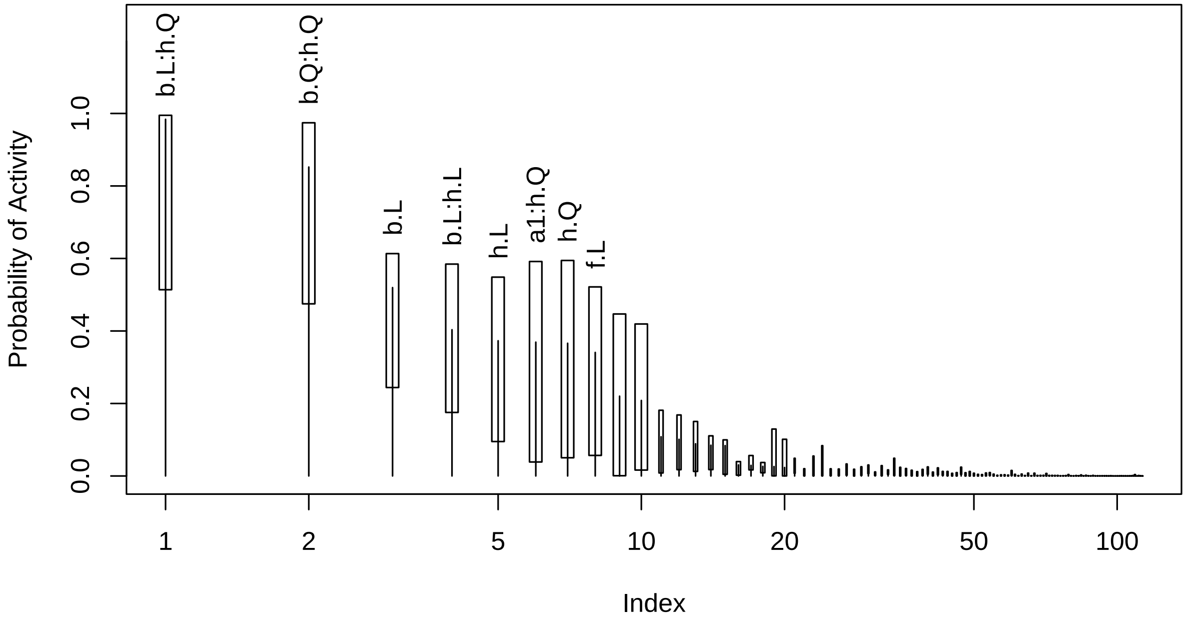

Figure 1: *Glucose Data: Marginal probability of activity.*

patterns between effects, it is effective at identifying different subsets that fit the data well. It is also applicable to regular fractional factorial designs.

Bayesian methods for subset selection offer several advantages over other approaches: The assigning of posterior probability to different subsets of active effects provides a way to characterize uncertainty about effect activity. Prior distributions can incorporate principles of effect dependence, such as effect heredity. The identification of promising models via Bayesian stochastic search techniques is faster than all subsets searches, and more comprehensive than stepwise methods.

The motivating example has illustrated the challenges of subset selection for screening experiments and the results of a Bayesian analysis. The remainder of the chapter provides details of the Bayesian approach. This section provides a brief outline of the main issues covered.

Some familiarity with Bayesian methods is assumed. Lee (1997) provides a good introduction without assuming too much advanced statistical background. Chapter 3 of Zellner (1987) provides detailed background on Bayesian multiple regression modeling. Bayesian simulation techniques, namely Markov chain Monte Carlo (MCMC), are overviewed in Chapter 11 of Gelman, Carlin, Stern & Rubin (1995).

Central to Bayesian approaches is the treatment of model parameters, such as a vector of regression coefficients $\boldsymbol{\beta}$, as random variables. Uncertainty and expert knowledge about these parameters are expressed via a prior distribution. The observed data gives rise to a likelihood for the parameters. The likelihood and prior distribution are combined to give a posterior distribution for the parameters. In subset selection for linear regression, the model is extended to include not only regression coefficient vector $\boldsymbol{\beta}$ and residual standard deviation $\sigma$, but also a vector $\boldsymbol{\delta}$ of binary indicators specifying whether each effect is active or inactive. That is $\boldsymbol{\delta}$ identifies a subset of active effects. Interest focuses on prior and posterior distributions for $\boldsymbol{\beta}, \sigma, \boldsymbol{\delta}$.

The Bayesian approach to subset selection is outlined in Sections 2-4. Section 2 outlines the mathematical ingredients of the analysis: A probability model for the data, prior distributions for the parameters of the model $(\boldsymbol{\beta}, \sigma, \boldsymbol{\delta})$, and the resultant posterior distribution.

For a particular data set, subsets with high posterior probability must be identified. This can be a computational challenge: with $p$ possible effects, there are as many as $2^p$ different subsets. To identify promising subsets, MCMC methods for simulating from the posterior distribution on subsets may be used as a stochastic search. Section 3 outlines efficient techniques for exploring the subset space with MCMC methods.

Section 4 reviews simple, semi-automatic methods of choosing the hyperparameters of a prior distribution, and adds some new insights into the choice of hyperparameters for a prior on regression coefficient vector $\boldsymbol{\beta}$.

The glucose experiment and a simulated data set are used in Section 5 to illustrate the application of the Bayesian subset selection technique.

A promising new use of prior distributions for subset selection is in optimality criteria for the construction of designs that allow model discrimination. This technique is discussed in Section 6. The paper concludes with a discussion, including possible extensions of the techniques to generalized linear models.

# 2 Model formulation

This section reviews the linear regression model, including an augmented form that allows specification of the subset of active effects. Prior distributions are introduced and relevant posterior distributions given.

## 2.1 The linear regression model

The usual regression model,

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

is used, with $\mathbf{Y}$ a $n$-vector of responses, $X$ an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ a $p$-vector of regression coefficients and $\boldsymbol{\varepsilon}$ a $n$-vector of errors. The $\varepsilon_i, i = 1, \ldots, n$ are assumed to be independent and identically distributed $N(0, \sigma^2)$. The columns of $X$ are functions of one or more of the original factors, such as linear, quadratic, or interaction contrasts, or indicator variables.

This model is augmented by an unobserved $p$-vector $\boldsymbol{\delta}$. Each element $\delta_j$ $(j = 1, \ldots, p)$ takes the value 0 or 1, indicating whether the corresponding $\beta_j$ is active $(\delta_j = 1)$ or inactive $(\delta_j = 0)$. An inactive effect is a $\beta_j$ close to 0, and an active effect is a $\beta_j$ far from 0. The precise definition of “active” and “inactive” may vary according to the form of prior distribution specified. Under this formulation, the Bayesian subset selection problem becomes one of identifying a posterior distribution on $\boldsymbol{\delta}$.

## 2.2 Prior distributions for subset selection in regression

A Bayesian analysis proceeds by placing prior distributions on the regression coefficient vector $\boldsymbol{\beta}$, error standard deviation $\sigma$ and subset indicator $\boldsymbol{\delta}$. This section outlines in detail one form of prior distribution and discusses other approaches. Techniques for choosing hyperparameters of prior distributions, such as the mean of a prior distribution, are discussed later in Section 4.

A variety of prior distributions have been proposed for the subset selection problem. Most differ in terms of the distribution placed on $\boldsymbol{\beta}$ and on the subset indicator $\boldsymbol{\delta}$. One particular formulation, due to Box & Meyer (1986) and George & McCulloch (1997), is reviewed in detail here. Subsequent examples will use this particular formulation, although many arising issues will be general.

The joint prior distribution on $\boldsymbol{\beta}, \sigma, \boldsymbol{\delta}$ can be factored and subsequently simplified by assuming that the subset indicator $\boldsymbol{\delta}$ and error variance are independent.

$$p(\boldsymbol{\beta}, \sigma, \boldsymbol{\delta}) = p(\boldsymbol{\beta}|\sigma, \boldsymbol{\delta})p(\sigma, \boldsymbol{\delta}) = p(\boldsymbol{\beta}|\sigma, \boldsymbol{\delta})p(\sigma)p(\boldsymbol{\delta}) \tag{2}$$

The prior distributions $p(\boldsymbol{\beta}|\sigma, \boldsymbol{\delta})p(\sigma)$ and $p(\boldsymbol{\delta})$ are described in Sections 2.2.1 and 2.2.2, respectively.

### 2.2.1 Prior distribution on regression parameters $\boldsymbol{\beta}, \sigma$

Prior distributions are often chosen to simplify the form of the posterior distribution. The posterior distribution is proportional to the product of the likelihood and the prior distribution, so if the prior distribution is of the same form as the likelihood, this simplification occurs. Such a choice is referred to as the use of a conjugate prior distribution (see Lee (1997) for details). In the regression model, the likelihood for $\boldsymbol{\beta}, \sigma$ can be written as the product of a normal distribution on $\boldsymbol{\beta}$ and an inverse gamma distribution on $\sigma$. This form motivates the conjugate choice of a normal-inverse-gamma prior distribution on $\boldsymbol{\beta}, \sigma$. Additional details on this prior can be found in Zellner (1987).

The prior distribution used for error variance,

$$\sigma^2 \sim \text{Inverse Gamma}(\nu/2, \nu\lambda/2) \tag{3}$$

is equivalent to $\nu\lambda/\sigma^2 \sim \chi^2_\nu$. This prior distribution is identical to the likelihood from a data set with $\nu$ observations and sample variance $\lambda$.

A variety of prior distributions $p(\boldsymbol{\beta}|\sigma, \boldsymbol{\delta})$ for $\boldsymbol{\beta}$ been proposed. Here, the following formulation is used: Conditional on the subset $\boldsymbol{\delta}$ and error variance $\sigma^2$, the effects $\boldsymbol{\beta}$ have an independent normal prior distribution. The variance of $\beta_j$ will be large or small depending on $\delta_j$:

$$p(\beta_j|\delta_j, \sigma) = \begin{cases} N(0, (\tau_j\sigma)^2) & \text{if } \delta_j = 0 \\ N(0, (c_j\tau_j\sigma)^2) & \text{if } \delta_j = 1 \end{cases}. \tag{4}$$

This will be referred to as a “mixture of two normal distributions”. The hyperparameters $c_j, \tau_j$ are chosen to indicate magnitudes of inactive and active effects. Roughly speaking, active effect $\beta_j$ will be $c_j$ times larger than if it was inactive. Choosing $c_j \gg 1$ represents this belief. Section 4 suggests how hyperparameters $c_j$ and $\tau_j$ may be selected automatically.

Implicit in (4) is the assumption of a diagonal prior covariance matrix for $\boldsymbol{\beta}$. Other choices are explored in Chipman, George & McCulloch (2001) and Raftery, Madigan & Hoeting (1997), including a prior covariance proportional to $(X'X)^{-1}$.

Alternates to prior distribution (4) have been proposed. An important practical difference is the extent to which they allow analytic simplification of the posterior, discussed in Section 2.3. Two alternate formulations are:

- George & McCulloch (1993) choose a prior distribution for $\boldsymbol{\beta}$ that does not depend on $\sigma$:

$$p(\beta_j|\delta_j, \sigma) = p(\beta_j|\delta_j) = \begin{cases} N(0, \tau_j^2) & \text{if } \delta_j = 0 \\ N(0, c_j^2\tau_j^2) & \text{if } \delta_j = 1 \end{cases}. \tag{5}$$

- Raftery et al. (1997) and Box & Meyer (1993) use a prior distribution similar to (4), but when $\delta_j = 0$, all prior probability on $\beta_j$ is a point mass at 0. This is a limiting case of (4), with $c \to \infty, \tau \to 0, c\tau$ fixed. It can be represented as

$$p(\beta_j|\delta_j, \sigma) = \begin{cases} \Delta(\beta_j) & \text{if } \delta_j = 0 \\ N(0, (c_j\tau_j\sigma)^2) & \text{if } \delta_j = 1 \end{cases}, \tag{6}$$

  where the Dirac delta function $\Delta$ assigns probability 1 to the event $\beta_j = 0$. Formally, $\Delta(x)$ integrates to 1, and takes the value zero everywhere except at $x = 0$.

### 2.2.2 Prior distribution on subset indicator $\boldsymbol{\delta}$

A prior distribution must also be assigned to the subset indicator $\boldsymbol{\delta}$. Since $\boldsymbol{\delta}$ is a binary $p$-vector, there are $2^p$ possible subsets. The prior distribution on $\boldsymbol{\delta}$ must assign probability to each subset. An initially appealing choice is to make each of the $2^p$ subsets equally likely. This is equivalent to prior independence of all elements of $\boldsymbol{\delta}$, and $p(\delta_j = 0) = p(\delta_j = 1) = 0.5$. In a screening context, this is implausible under the following widely accepted principles:

1. **Effect Sparsity:** Only a small fraction of all possible effects are likely to be active. Thus the prior probability that $\delta_j = 1$ will be less than 0.5.
2. **Effect Hierarchy:** Lower order effects are more likely to be active than higher-order effects. For example, linear main effects are more likely to be active than quadratic main effects or interaction effects.
3. **Effect Heredity:** Subsets should obey heredity of active effects. For example, a subset with an active $AB$ interaction but no $A$ or $B$ main effects may not be acceptable. Nelder (1998) refers to this as the "marginality principle".

The first two points suggest that $P(\delta_j = 1)$ should be less than 0.5 and be smaller for higher order effects. The third suggests that $P(\boldsymbol{\delta})$ should incorporate effect dependencies. Chipman (1996) proposes such a structure. The probability that a given term is active or inactive depends on its "parent" terms, typically taken to be those terms of the next lowest order from which the given term may be formed. To illustrate, consider a simple example with three factors $(A, B, C)$ and a model with 3 linear main effects and 3 linear×linear interactions. For clarity, elements of $\boldsymbol{\delta}$ are indexed as $\delta_A, \delta_{AB}$,etc., and since all effects are linear or linear×linear interactions, the $L$ subscript is omitted from linear effect $A_L$. The prior distribution on $\boldsymbol{\delta}$ is

$$\begin{aligned} p(\boldsymbol{\delta}) &= p(\delta_A, \delta_B, \delta_C, \delta_{AB}, \delta_{AC}, \delta_{BC}) \\ &= p(\delta_A, \delta_B, \delta_C) p(\delta_{AB}, \delta_{AC}, \delta_{BC} | \delta_A, \delta_B, \delta_C) \qquad (7) \end{aligned}$$

The effect heredity principle motivates two simplifying assumptions. The first assumption is that the activity of equal order terms is independent, given the activity of lower order terms:

$$p(\boldsymbol{\delta}) = p(\delta_A)p(\delta_B)p(\delta_C)p(\delta_{AB}|\delta_A, \delta_B, \delta_C)p(\delta_{AC}|\delta_A, \delta_B, \delta_C)p(\delta_{BC}|\delta_A, \delta_B, \delta_C)$$

Second, the activity of an interaction is assumed to depend only on the activity of those terms from which it is formed:

$$p(\boldsymbol{\delta}) = p(\delta_A)p(\delta_B)p(\delta_C)p(\delta_{AB}|\delta_A, \delta_B)p(\delta_{AC}|\delta_A, \delta_C)p(\delta_{BC}|\delta_B, \delta_C)$$

The prior distribution is specified by choosing marginal probabilities that a main effect is active,

$$P(\delta_A = 1) = \pi \qquad (8)$$

and the conditional probability that an interaction is active,

$$P(\delta_{AB} = 1|\delta_A, \delta_B) = \begin{cases} \pi_0 & \text{if } (\delta_A, \delta_B) = (0,0) \\ \pi_1 & \text{if one of } \delta_A, \delta_B = 1 \\ \pi_2 & \text{if } (\delta_A, \delta_B) = (1,1) \end{cases} . \tag{9}$$

Although in principle four probabilities could be specified in (9), the circumstances under which $(\delta_A, \delta_B) = (0,1)$ and (1,0) can be distinguished are uncommon. In some applications, $\pi, \pi_0, \pi_1, \pi_2$ may vary for (say) effects associated with $A$ and with $B$. To keep notation simple, this straightforward generalization is not discussed further.

Effect sparsity and effect hierarchy will be represented by the choice of hyperparameters $\pi, \pi_0, \pi_1, \pi_2$. Typically $\pi_0 < \pi_1 < \pi_2 < \pi < 0.5$. Section 2.3 provides details on the selection of these hyperparameters.

Choosing $\pi_0 = \pi_1 = 0, \pi_2 > 0$ in (9) allows an interaction to be active only if both corresponding main effects are active (referred to as *strong heredity*). Choosing $\pi_0 = 0, \pi_1 > 0, \pi_2 > 0$ allows an interaction to be active if one or more of its parents are active (*weak heredity*). Models obeying strong heredity are usually easier to interpret, while weak heredity may help the stochastic search explore the model space. Peixoto (1990) and Nelder (1998) argue in favor of strong heredity, since models identified are invariant to linear transformations of the factor codings. Chipman, Hamada & Wu (1997) suggest that in exploratory stages, it may be desirable to relax the restrictions of strong heredity. They instead use the weak-heredity prior distribution, with $\pi_1 < \pi_2$ to indicate a preference for strong-heredity models. A different choice of parameters in (9) is given by Box & Meyer (1993), in which $\pi_0 = 0, \pi_1 = 0, \pi_2 = 1$. An interaction would be active only if all its main effect parents are active, in which case it is forced to be active. This will be referred to as an "effect forcing" prior distribution. Effect forcing greatly reduces the number of models under consideration, since activity of interactions is automatically determined by the activity of main effects.

Generalizations to quadratic effects and polynomial interactions are possible. Here, linear and quadratic contrasts will be denoted $B_L$ and $B_Q$. A linear×quadratic interaction will be denoted $B_L H_Q$. A higher order term such as $A_L B_Q$ could have parents $A_L B_L$ and $B_Q$, or have parents $A_L$ and $B_Q$. Chipman (1996) and Chipman et al. (1997) adopt the convention that a term's parents are terms of the next lowest order that, when multiplied by a main effect, would generate the term. Thus $A_L B_Q$ would have parents $A_L B_L$ and $B_Q$. Figure 2 illustrates this relationship. Each term that is a function of two factors will have two parents. Probabilities such as (9) would be specified for higher order polynomials, and for polynomial terms in a single variable, one might specify

$$P(\delta_{A_Q} = 1|\delta_{A_L}) = \begin{cases} \pi_3 & \text{if } \delta_{A_L} = 0 \\ \pi_4 & \text{if } \delta_{A_L} = 1 \end{cases} . \tag{10}$$

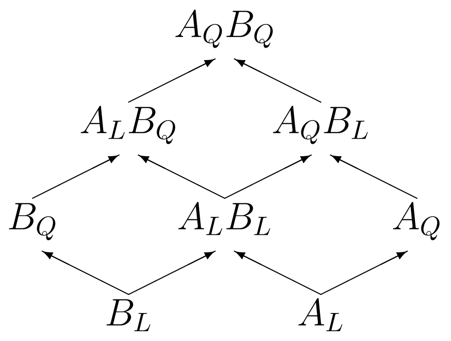

Figure 2: *Ordering and inheritance relations among polynomial interactions*

The hyperparameters $\pi_3, \pi_4$ would often be chosen as $\pi_3 = \pi_0, \pi_4 = \pi_2$.

If some factors are categorical with more than 2 levels, the corresponding effects will be estimated via coding of indicator variables in the regression matrix. Chipman (1996) suggests a prior distribution in which a single indicator $\delta_F$ controls the activity of all $l$ effects $\beta_{F_1}, \beta_{F_2}, \ldots, \beta_{F_l}$ of a factor $F$ with $l+1$ levels. This *effect grouping* allows either all or none of $\beta_{F_1}, \ldots, \beta_{F_l}$ to be active.

## 2.3 The posterior distribution on subset indicator $\boldsymbol{\delta}$

The joint posterior distribution of $\boldsymbol{\beta}, \sigma, \boldsymbol{\delta}$ will proportional to the product of a likelihood from the linear model (1) and the prior distribution (2):

$$p(\boldsymbol{\beta}, \sigma, \boldsymbol{\delta} | \mathbf{Y}) \propto p(\mathbf{Y} | \boldsymbol{\beta}, \sigma, \boldsymbol{\delta}) p(\boldsymbol{\beta} | \sigma, \boldsymbol{\delta}) p(\sigma) p(\boldsymbol{\delta}) \tag{11}$$

Of primary interest in the subset selection problem is the marginal posterior distribution of the subset indicator $\boldsymbol{\delta}$:

$$\begin{aligned} p(\boldsymbol{\delta} | \mathbf{Y}) &\propto \iint p(\mathbf{Y} | \boldsymbol{\beta}, \sigma, \boldsymbol{\delta}) p(\boldsymbol{\beta} | \sigma, \boldsymbol{\delta}) p(\sigma) p(\boldsymbol{\delta}) d\boldsymbol{\beta} d\sigma \\ &= p(\boldsymbol{\delta}) \iint p(\mathbf{Y} | \boldsymbol{\beta}, \sigma, \boldsymbol{\delta}) p(\boldsymbol{\beta} | \sigma, \boldsymbol{\delta}) d\boldsymbol{\beta} p(\sigma) d\sigma \\ &= p(\boldsymbol{\delta}) p(\mathbf{Y} | \boldsymbol{\delta}). \end{aligned} \tag{12}$$

The integral in (12) is either evaluated by MCMC methods (described in Section 3.1) or analytically. The result of integration, $p(\mathbf{Y}|\boldsymbol{\delta})$, will be referred to as the marginal likelihood of $\boldsymbol{\delta}$, since it is the likelihood of $(\boldsymbol{\beta}, \sigma, \boldsymbol{\delta})$ integrated over the prior distribution of $(\boldsymbol{\beta}, \sigma)$. Conjugate prior distributions (4) or (6) allow analytic integration. The nonconjugate prior distribution of George & McCulloch (1993) requires MCMC integration.

For prior distribution (3) and (4), the marginal likelihood of $\boldsymbol{\delta}$ in (12) is given by (see George

& McCulloch (1997)):

$$p(\mathbf{Y}|\boldsymbol{\delta}) \propto |\tilde{X}'\tilde{X}|^{-1/2}|D_{\boldsymbol{\delta}}|^{-1}(\lambda\nu + S^2_{\boldsymbol{\delta}})^{(n+\nu)/2}, \tag{13}$$

where $D_{\boldsymbol{\delta}}$ is diagonal with $j$th element $\tau_j(1-\delta_j) + c_j\tau_j\delta_j$,

$$S^2_{\boldsymbol{\delta}} = \tilde{\mathbf{Y}}'\tilde{\mathbf{Y}} - \tilde{\mathbf{Y}}'\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{\mathbf{Y}},$$

$$\tilde{\mathbf{Y}} = \begin{bmatrix} \mathbf{Y} \\ 0 \end{bmatrix} \qquad \text{and} \qquad \tilde{X} = \begin{bmatrix} X \\ D^{-1}_{\boldsymbol{\delta}} \end{bmatrix}.$$

The posterior on subset indicator $\boldsymbol{\delta}$ is then given by

$$p(\boldsymbol{\delta}|\mathbf{Y}) \propto p(\boldsymbol{\delta})p(\mathbf{Y}|\boldsymbol{\delta}) \equiv g(\boldsymbol{\delta}) \tag{14}$$

George & McCulloch (1997) also give analytic results for a point mass mixture prior distribution.

Whether the marginal posterior distribution on subsets is calculated analytically (as in (13) and (14)) or via MCMC, there still remains the challenge of identifying subsets of active effects ($\boldsymbol{\delta}$ values) that have high posterior probability. The form of (13) suggests a similarity with least squares regression: the $S^2_{\boldsymbol{\delta}}$ term acts as a residual sum of squares for a model relating $\tilde{\mathbf{Y}}$ to $\tilde{X}$. Thus in principle, efficient all-subsets search algorithms (Furnival & Wilson 1974) could be used to identify subsets with high marginal likelihood (13). The marginal likelihoods would then be multiplied by prior probability $p(\boldsymbol{\delta})$ to obtain posterior (14). The limitations of this approach have already been illustrated in Section 1: slow computation in large problems, and the vast majority of models with high $R^2$ will not obey strong or weak heredity.

For these reasons, stochastic search methods based on MCMC techniques are a popular alternative. These are discussed in the next section.

# 3 Efficient stochastic search for active subsets

The focus of this section is MCMC methods for sampling from $p(\boldsymbol{\delta}|\mathbf{Y})$, the posterior distribution of $\boldsymbol{\delta}$. In some cases such as prior distribution (5), MCMC is also used to simulate from $p(\boldsymbol{\beta}, \sigma, \boldsymbol{\delta}|\mathbf{Y})$.

The large number of models makes it impractical to evaluate the posterior probability (14) for all possible subsets of active effects. Instead MCMC methods (either the Gibbs sampler or the Metropolis-Hastings algorithm) are used to draw samples from the posterior distribution. George & McCulloch (1997) discuss both the Metropolis-Hastings and Gibbs sampling algorithms for subset selection. In the context of subset selection, MCMC algorithms may be thought of as a stochastic search.

### 3.1 MCMC sampling for subset indicator $\boldsymbol{\delta}$

The Gibbs sampler will be used to simulate from $p(\boldsymbol{\delta}|\mathbf{Y})$. The algorithm starts with initial values of all parameters and then repeatedly draws each parameter conditional on all the others and the data. The steps below would generate $K$ draws $\boldsymbol{\delta}^1, \boldsymbol{\delta}^2, \ldots, \boldsymbol{\delta}^K$ that converge to the posterior distribution of $\boldsymbol{\delta}$:

1. Choose an initial value, $\boldsymbol{\delta}^0 = (\delta_1^0, \delta_2^0, \ldots, \delta_p^0)$
2. For $i = 1, \ldots, K$:
   (a) For $j = 1, \ldots, p$:
      - Draw $\delta_j^i$ from $p(\delta_j^i|\delta_1^i, \ldots, \delta_j^i, \delta_{j+1}^{i-1}, \ldots \delta_p^{i-1}, \mathbf{Y})$.

Each draw is from a Bernoulli distribution. In drawing $\delta_j^i$, the $j$th component of $\boldsymbol{\delta}^i$, we condition on the most recently drawn values of all other components of $\boldsymbol{\delta}$. All values in the generated sequence $\boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^K$ will be treated as draws from the posterior distribution of $\boldsymbol{\delta}$.

For the nonconjugate prior distribution (5), George & McCulloch (1993) and Chipman et al. (1997) used the Gibbs sampler to simulate the joint posterior distribution of $(\boldsymbol{\beta}, \sigma, \boldsymbol{\delta})$. That is, the above algorithm would have a step 2(b) to draw from $p(\boldsymbol{\beta}|\boldsymbol{\delta}, \sigma, \mathbf{Y})$ and a step 2(c) to draw from $p(\sigma|\boldsymbol{\beta}, \mathbf{Y})$.

### 3.2 Estimation of posterior probability on $\boldsymbol{\delta}$ using MCMC output

In both the conjugate and nonconjugate cases just described, the most natural estimator of the posterior probability of a subset $\boldsymbol{\delta}'$ is the observed relative frequency of $\boldsymbol{\delta}'$ among the $K$ sampled subsets $\mathcal{S} = \boldsymbol{\delta}^1, \boldsymbol{\delta}^2, \ldots, \boldsymbol{\delta}^K$,

$$\widehat{p}(\boldsymbol{\delta}'|\mathbf{Y}) = \frac{\sum_{i=1}^K I(\boldsymbol{\delta}^i = \boldsymbol{\delta}')}{K}. \tag{15}$$

Here, the indicator function $I(\cdot)$ is 1 whenever its argument is true, and zero otherwise.

There are some problems with relative frequency estimate (15). First, it is prone to variability in the MCMC sample. Second, any model that is not in $\mathcal{S}$ has an estimated posterior probability of 0. Third, if the starting value $\boldsymbol{\delta}^0$ has very low posterior probability, it may take the chain a large number of steps to move to $\boldsymbol{\delta}$ values that have high posterior probability. These initial "burn-in" values of $\boldsymbol{\delta}$ would have larger estimates of posterior probability in (15) than their actual posterior probability. That is, the estimate $\hat{p}(\boldsymbol{\delta}|\mathbf{Y})$ will be biased because of the burn-in. For example, with the simulated data (described later in Section 4.2), the first 100 draws of $\boldsymbol{\delta}$ have almost no posterior probability. In a run of $K = 1000$ steps, the relative frequency estimate (15) of posterior probability of these 100 draws would be 1/10, while the actual posterior probability is nearly zero.

In conjugate settings such as (4) or (6), a better estimate of posterior probability $p(\boldsymbol{\delta}|\mathbf{Y})$ is available. Instead of the relative frequency estimate (15), the analytic expression for posterior probability (14) is used:

$$p(\boldsymbol{\delta}'|\mathbf{Y}) = Cg(\boldsymbol{\delta}), \tag{16}$$

provided that the normalizing constant $C$ can be estimated from the MCMC draws $\mathcal{S}$. Two methods for estimating $C$ are discussed below. The analytic estimate of posterior probability (16) will always be used in this paper, rather than the relative frequency estimate (15). The analytic estimate solves the problems of sampling variation and bias in estimating posterior probability due to burn-in described above.

The first approach to estimating normalizing constant $C$ is to renormalize the probabilities for all unique subsets in the sampled set $\mathcal{S}$ so that they sum to 1. Let $\mathcal{U}$ be the set of unique $\boldsymbol{\delta}$ values in $\mathcal{S}$. The constant $C$ is then estimated by

$$\widehat{C} = \frac{1}{\sum_{i|\boldsymbol{\delta}^i \in \mathcal{U}} g(\boldsymbol{\delta}^i)}. \tag{17}$$

The estimate (17) will be biased upwards when all unvisited subsets are assigned probability zero. A better estimate of $C$ can be obtained by a capture-recapture approach, as in George & McCulloch (1997). Let the initial "capture set" $\mathcal{A}$ be a collection of $\boldsymbol{\delta}$ values identified before a run of the MCMC search. That is, each element in the set $\mathcal{A}$ is a particular subset. The "recapture" estimate of the probability of $\mathcal{A}$ is the relative frequency as in (15). The analytic expression (16) for posterior probability is also available, and contains the unknown $C$. Let $g(\mathcal{A}) = \sum_{\boldsymbol{\delta} \in \mathcal{A}} g(\boldsymbol{\delta})$ so that $p(\mathcal{A}|\mathbf{Y}) = Cg(\mathcal{A})$.

By equating the two estimates,

$$Cg(\mathcal{A}) = \sum_{i=1}^{K} I(\boldsymbol{\delta}^i \in \mathcal{A})/K$$

and solving for $C$, a consistent estimator of $C$ is obtained:

$$\widehat{C} = \frac{1}{g(\mathcal{A})K} \sum_{k=1}^{K} I_{\mathcal{A}}(\boldsymbol{\delta}^k), \tag{18}$$

This technique is illustrated in Section 5.1.

Having shown that the observed frequency distribution of the sampler is useful in the estimation of normalizing constant $C$, it is interesting to note that the frequencies are otherwise unused. As long as the posterior probability of a subset is available as an analytic expression, it is preferred over the frequencies. In such a context, the main goal of the sampler is to visit as many high probability models as possible. Visits after the first add no value, since the model has already been identified for analytic evaluation of $p(\boldsymbol{\delta}|\mathbf{Y})$.

However posterior probabilities are estimated, arbitrary quantities of interest can be estimated by averaging their value over posterior distribution (17) or (15). For example, the marginal probability of activity for a main effect $A$ could be calculated as the expected value of $\delta_A$, where $\delta_A$ is the component of $\boldsymbol{\delta}$ corresponding to main effect $A$. Based on the unique draws $\mathcal{U}$, the estimated posterior marginal probability would be

$$\Pr(\delta_A = 1|\mathbf{Y}) \approx \sum_{i \in \mathcal{U}} \delta_A^i \widehat{C} g(\boldsymbol{\delta}^i) \tag{19}$$

Either (17) or (18) could be used to estimate $\widehat{C}$ in (19). (17) is preferred, since it produces estimates between 0 and 1. (18) would produce estimates less than 1, since $\sum_{i \in \mathcal{U}} \widehat{C} g(\boldsymbol{\delta}^i) < 1$ when $\widehat{C}$ is from (18).

## 3.3 The impact of prior distributions on computation

The prior distributions discussed in Section 2.2 impact both the ease of implementation of the MCMC algorithm, and its speed of execution. Specific issues include the number of linear algebra operations and the rate at which stochastic search methods can explore the space of all subsets.

The point mass prior (6) on $\boldsymbol{\beta}$ reduces computation by dropping columns from the $X$ matrix. With this prior, $\delta_j = 0$ implies $\beta_j = 0$, and the corresponding column is dropped. When a mixture of two normal distributions (4) is used, the $X$ matrix is always of dimension $p$.

Computations are also more efficient if MCMC can sample directly from the marginal posterior distribution $p(\boldsymbol{\delta}|\mathbf{Y})$, rather than from the joint posterior distribution $p(\boldsymbol{\delta}, \boldsymbol{\beta}, \sigma|\mathbf{Y})$. This efficiency occurs because fewer variables are being sampled. As mentioned at the end of Section 3.1, the marginal posterior distribution $p(\boldsymbol{\delta}|\mathbf{Y})$ is available in closed form when conjugate prior distributions on $\boldsymbol{\beta}$, (4) or (6), are used.

The prior distribution on $\boldsymbol{\delta}$ affects the number of possible subsets of active effects to be searched. Figure 3 plots the number of possible subsets against the number of factors when linear main effects and linear×linear interactions are considered. When there are $m$ factors, there will be $p = m + \binom{m}{2} = (3m + m^2)/2$ effects under consideration. The number of possible subsets will be $2^p$ if all possible subsets of active effects are considered. Hence a log base 2 scale is used on the vertical axis. Strong and weak heredity reduce somewhat the number of subsets. The effect forcing of Box & Meyer (1993) yields far fewer subsets, since activity of interactions is automatically determined by activity of main effects. With effect forcing, posterior probability can easily be calculated for every possible subset, rather than via stochastic search methods.

Prior distributions that enforce heredity can also affect the manner in which stochastic search algorithms move around the subset space. The Gibbs sampling algorithm in Section 3.1 updates one element of $\boldsymbol{\delta}$ at a time, conditional on the values of all other elements. That is, the algorithm randomly adds or drops a single effects from the subset in each step. This is a common approach of many stochastic search algorithms. Strong heredity restricts the number

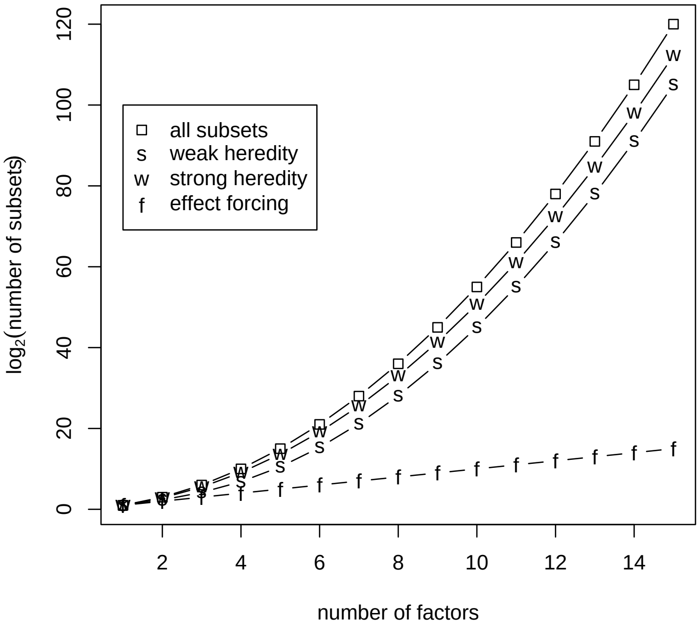

Figure 3: *Number of possible subsets (*$\log_2$ *scale) against number of factors for models with two-way interactions and main effects only.*

of paths between subsets. For example, to move from the subset $\{A, B, C, AB, AC\}$ to the subset $\{A\}$, it would be necessary to drop $AB$ before dropping $B$ and $AC$ before dropping $C$. Weak heredity, which allows subsets like $\{A, C, AB, AC\}$, provides more paths between subsets, enabling the stochastic search to move more freely around the subset space.

Prior distributions that enforce effect grouping, such as those for indicator variables also impact computation. The Gibbs sampler usually draws $\boldsymbol{\delta}$ elements one at a time. Changing a single element $\delta_j$ affects only one $\boldsymbol{\beta}$ element, $\beta_j$. With grouped effects, one $\boldsymbol{\delta}$ element implies a change to multiple $\boldsymbol{\beta}$ elements. Since updating formulae are used in calculating how changes to $\boldsymbol{\beta}$ affect the posterior distribution, a simultaneous change to multiple $\boldsymbol{\beta}$ may entail more complicated updates.

# 4 Selection of hyperparameters of the prior distribution

This section outlines choices for hyperparameters of the prior distribution (3) and (4), with an emphasis on automatic methods that use simple summaries of the observed data. Many of these suggestions have been made in Chipman et al. (1997) and Chipman (1998). A few relating to choice of $c$ and $\tau$ are new.

The term "hyperparameter" will be used throughout this section to refer to the hyperparameter of a prior distribution.

## 4.1 Hyperparameters for the $\sigma$ prior distribution

I propose that the hyperparameters for the prior distribution (3) on $\sigma^2$ are chosen so that the mean and 99th quantile of the distribution are consistent with the observed values of the response. The prior expected value of $\sigma^2$ is

$$\mathrm{E}(\sigma^2) = \frac{\lambda\nu}{\nu - 2} \qquad \text{for } \nu > 2,$$

suggesting that $\lambda$ be chosen near the expected residual variance. In the absence of expert knowledge, some fraction of the sample variance of the response, $s^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2/(n-1)$ could be used to choose $\lambda$. Chipman et al. (1997) propose

$$\lambda = s^2/25 = \left[\sum_{i=1}^{n}(y_i - \overline{y})^2/(n-1)\right]/25.$$

This represents the belief that the residual standard deviation will be roughly 1/5 the standard deviation of the response when there is no model.

As was observed after (3) in Section 2.2.1, the hyperparameter $\nu$ can be thought of as the amount of information about $\sigma$ arising from a sample with $\nu$ observations and sample variance $\lambda$. The parameter $\nu$ controls the length of the right tail of prior distribution (3) for $\sigma$. Larger

| $\nu$ | mean | 0.01 | 0.1 | 0.5 | 0.9 | 0.99 |
|---|---|---|---|---|---|---|
| 1 | – | 0.15 | 0.37 | 2.2 | 63.33 | 6365 |
| 2 | – | 0.22 | 0.43 | 1.44 | 9.49 | 99.50 |
| 3 | 3 | 0.26 | 0.48 | 1.27 | 5.13 | 26.13 |
| 4 | 2 | 0.30 | 0.51 | 1.19 | 3.76 | 13.46 |
| 5 | 1.67 | 0.33 | 0.54 | 1.15 | 3.10 | 9.02 |
| 6 | 1.5 | 0.36 | 0.56 | 1.12 | 2.72 | 6.88 |
| 7 | 1.4 | 0.38 | 0.58 | 1.10 | 2.47 | 5.65 |
| 8 | 1.33 | 0.40 | 0.60 | 1.09 | 2.29 | 4.86 |
| 9 | 1.29 | 0.42 | 0.61 | 1.08 | 2.16 | 4.31 |
| 10 | 1.25 | 0.43 | 0.63 | 1.07 | 2.06 | 3.91 |

Table 5: *Quantiles of an Inverse Gamma distribution with $\lambda = 1$. For other $\lambda$, multiply the table entries by $\lambda$ to obtain the appropriate quantile.*

values of $\nu$ imply a prior distribution that is more tightly centered around $\lambda$. Table 5 gives various quantiles for an inverse gamma with $\lambda = 1$, and indicates that the distribution has quite a long tail for $\nu \leq 5$. A sufficiently diffuse prior distribution may be selected by choosing $\nu$ so that the upper tail (say the 99th percentile) is not far below $s^2$. Choosing $\nu = 5$ would place the 99th percentile of the prior distribution at $9.02\lambda$, for example. Combining this with $\lambda = s^2/25$ gives the 99th prior quantile of $\sigma$ as $9.02s^2/25 = 0.36s^2$. Smaller $\nu$ are possible (for example values of $\nu = 1.5$ were used in Chipman et al. (1997)), although they can lead to unreasonably long tails, because the variance of an inverse gamma distribution is not defined for $\nu \leq 4$. In general, one would choose

$$\nu = 5 \quad \text{or from Table 5.}$$

## 4.2 Hyperparameters for the $\boldsymbol{\beta}$ prior distribution

Prior distribution (4) for $\boldsymbol{\beta}$ is defined by hyperparameters $c_j$ and $\tau_j$. In choosing these hyperparameters, it is helpful to recall from (4) that the coefficient $\beta_j$ associated with an inactive contrast has standard deviation $\sigma\tau_j$ and if the contrast is instead active, $\beta_j$ has a standard deviation that is $c_j$ times larger.

Box & Meyer (1986) suggest $c = 10$, separating large and small coefficients by an order of magnitude. George & McCulloch (1997) suggest the following technique for choice of $\tau_j$: For an inactive contrast, even a large change in the contrast value $X$ should produce a small change in the mean of response $Y$. Such a small change in the response (say $\Delta Y$) could be considered to be equal to residual standard deviation $\sigma$. A large change in $X$ (say $\Delta X$) will be taken to be the $\max(X) - \min(X)$ over the contrast values set in the design. A small coefficient $\beta_j$ has standard deviation $\sigma\tau_j$ and will lie within $0 \pm 3\sigma\tau_j$ with very high probability. Thus, when $X$ changes by a large amount $\Delta X$, the mean of $Y$ is unlikely to change by more than $3\sigma\tau_j\Delta X$. Solving $\sigma = 3\sigma\tau_j\Delta X$ gives $\tau_j = 1/3\Delta X = 1/3(\max(X) - \min(X))$. In two-level designs, in which contrasts are coded as ±1, $\Delta X = 2$. In summary, the default choice of hyperparameter values is

$$c_j = 10, \tau_j = \frac{1}{3 \times (\max(X_j) - \min(X_j))}. \tag{20}$$

An alternate choice, (21), is discussed later in this section.

The use of minimum and maximum values for each contrast makes the method invariant to rescalings of the contrasts. In the glucose example, contrasts are coded with quite different ranges. This will not impact the analysis because the definition of large and small effects is adjusted accordingly.

The subset selection procedure can be sensitive to the choice of $\tau_j$ (see Chipman et al. (1997) and George & McCulloch (1993)). Equation (20) captures the relative magnitudes of the $\tau_j$ for different variables, but the overall magnitude may need tuning. Box & Meyer (1993) and Chipman et al. (1997) propose methods for tuning based on runs of the search algorithm. A faster alternative (Chipman 1998) based on predictive distributions is discussed here. For any given subset $\boldsymbol{\delta}$, the posterior mean of $\mathbf{Y}$ for a given $X$ may be calculated, using the posterior mean of $\boldsymbol{\beta}$. The magnitude of $\tau_j$ will determine the degree of shrinkage for coefficient $\beta_j$, in a manner similar to ridge regression. A simple way to assess the impact of $\tau_j$ is to see how the predictions vary for a range of values $r\tau_j$, $r \in (1/10, 10)$, for a single given model. A good $\tau_j$ value would be the smallest value not shrinking posterior predictions too far from the least squares predictions.

The posterior mean for $\boldsymbol{\beta}$ conditional on a particular subset $\boldsymbol{\delta}$ is obtained by integration of $p(\boldsymbol{\beta}, \sigma | \mathbf{Y}, \boldsymbol{\delta})$ with respect to $\sigma$ to obtain $p(\boldsymbol{\beta} | \mathbf{Y}, \boldsymbol{\delta})$, and calculating an expectation of $\boldsymbol{\beta}$ with respect to this distribution. The resultant posterior mean is

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\delta}} = (X'X + D_{\boldsymbol{\delta}}^{-2})^{-1} X'\mathbf{Y}$$

where $D_{\boldsymbol{\delta}}$ is diagonal with elements $\tau_j(1 - \delta_j) + \tau_j c_j \delta_j$. See George & McCulloch (1997) for details.

To illustrate this idea, two examples are considered. First, a simulated example, and then the glucose data.

Hamada & Wu (1992) and Chipman et al. (1997) considered a simulated screening experiment with a 12-run Plackett-Burman design, given in Table 6. The 11 factors are labeled $A - K$, each

Table 6: 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 |
| − | + | − | + | + | − | + | + | + | − | − | -0.955 |
| + | − | + | + | − | + | + | + | − | − | − | 0.644 |
| − | + | + | − | + | + | + | − | − | − | + | -5.025 |
| − | − | − | − | − | − | − | − | − | − | − | 3.060 |

set at 2 levels (±1). The response was simulated from the true model

$$Y = A + 2AB + 2AC + \varepsilon, \qquad \varepsilon \sim N(0, \sigma = 0.25).$$

That is, factor $A$ has an active main effect and there are active interactions between $A$ and $B$ and between $A$ and $C$, while the remaining factors $D - K$ are inactive.

Figure 4 plots predicted values (that is, the posterior mean of $\mathbf{Y}$) in the simulated example. There are 12 design points and the default choice of $\tau$ is multiplied by $r \in (1/10, 10)$. The subset used (A, B, C, AB, AC) was identified by first finding the subset $A, AB, AC$ via stepwise regression. Main effects for $B$ and $C$ were subsequently included so that the subset obeys strong heredity. The "1" value on the horizontal axis is the default choice (20) for $\tau$. Both the observed $Y_i$ values (•) and predictions based on least squares estimate $\widehat{\boldsymbol{\beta}}$ (◦) are shown on the right side of the plot. In this case, the default seems quite reasonable, as any smaller multiples would shrink the posterior mean away from the data (•) and the least squares estimates (◦).

Figure 5 (a) gives a similar plot for the glucose data, using a subset of active effects $B_L, H_L, B_Q, H_Q, B_LH_L, B_LH_Q, B_QH_L, B_QH_Q$. Close inspection of this plot reveals a problem: The posterior mean of $Y_i$ converges to the *data values* (•) rather than the least squares estimate (◦). This is somewhat unexpected: one might suppose that as $\tau_j$ increased, the posterior means would approach the least squares estimates. The posterior means converge instead to the data

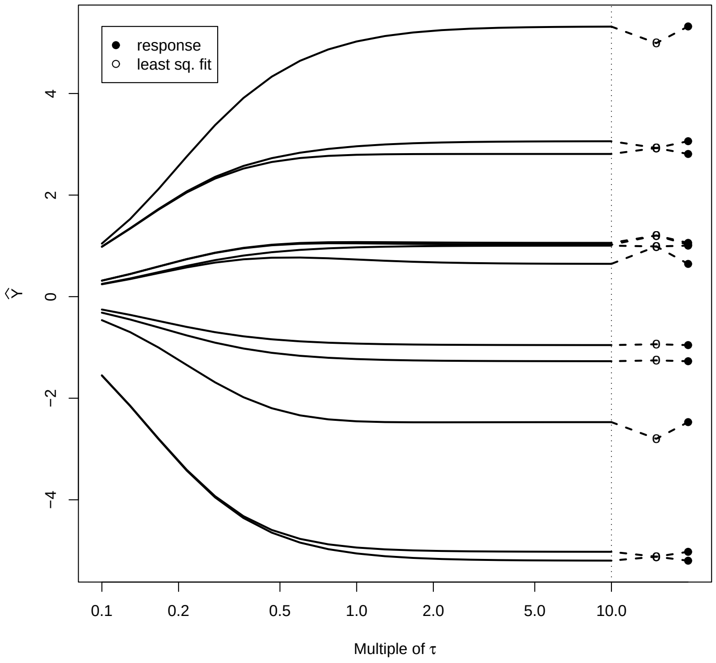

Figure 4: *Predicted response under original design for various multiples of hyperparameter* $\tau$

because, in addition to the intercept and 8 effects ($B_L, \ldots, B_Q H_Q$), there are $113 - 8 = 105$ other "inactive" effects included in the model. These are heavily shrunk towards zero, but not set exactly to 0, since prior (4) specifies that an inactive effect has prior standard deviation $\tau\sigma$. Some of the residual variation not captured by the eight active effects is absorbed by the 105 inactive effects, rather than being captured in the error term $\varepsilon$.

The tendency of inactive effects to capture residual variation suggests an alternate choice for $c_j$ and $\tau_j$ in experiments with a very large number of candidate effects: Reduce the prior magnitude of an inactive effect. This can be accomplished by multiplying $c_j$ in (20) by 10, and dividing $\tau_j$ in (20) by 10, giving alternate default choices of $c_j, \tau_j$:

$$c_j = 100, \tau_j = \frac{1}{30 \times \text{range}(X_j)}. \tag{21}$$

Since an active effect has prior standard deviation $\sigma c_j \tau_j$, this has no effect on the active effects. The prior standard deviation $\sigma\tau_j$ of a small effect has been shrunk by a factor of 10, approaching the point mass prior distribution of Raftery et al. (1997) and Box & Meyer (1993).

Considering multiples of the default $\tau_j$ in (21) in the glucose example produces Figure 5 (b). With a multiplier of 1-2, the posterior means first approach the least squares estimates. For $\tau$ in this range, the inactive effects are still shrunk quite close to zero and unable to absorb residual errors. Only when the multiplier of $\tau$ is quite large (such as a factor of 10 times the default) do

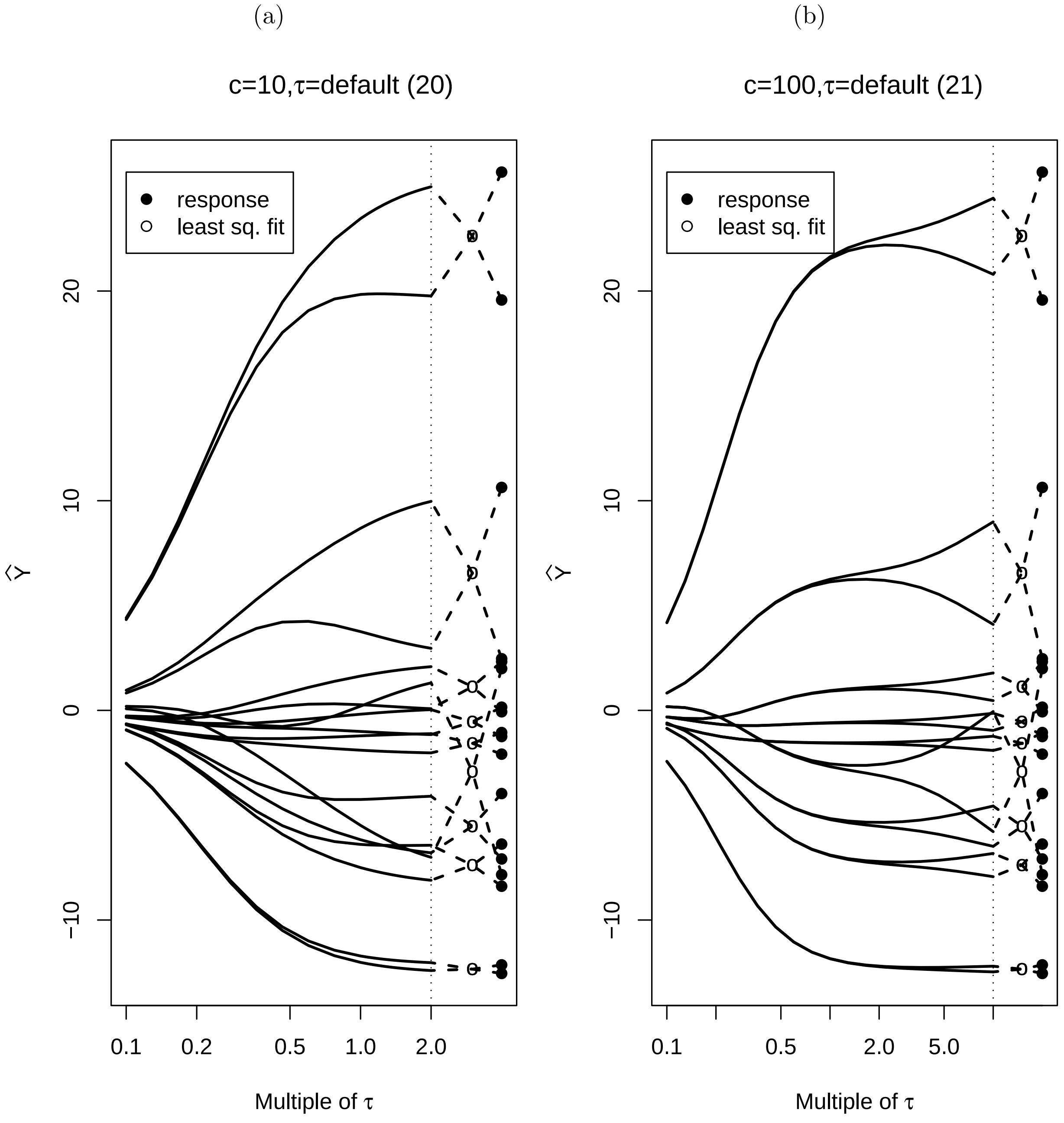

Figure 5: *Glucose data: Predicted response under original design for various multiples of hyperparameter* $\tau$*. Shrinkage relative to (20) and (21) are considered in panels (a) and (b) respectively. The model used has effects* $B_L, H_L, B_Q, H_Q, B_LH_L, B_LH_Q, B_QH_L, B_QH_Q$.

the fitted values converge to the data points.

### 4.3 Hyperparameters for the prior distribution on subset indicator $\boldsymbol{\delta}$

Box & Meyer (1986) examined several published analyses of experiments and concluded that between 10% and 30% of main effects were identified as active, with an average of 20%. Similar arguments may be made for screening experiments, with some modification for interactions and higher-order effects.

The suggestions made here utilize calculations for the expected number of active effects. Such expectations should be easier to specify than prior probabilities.

Before developing these expectations, a simplified method of choosing the hyperparameters is reviewed (Bingham & Chipman 2002). They assume that the probability of an active interaction is proportional to $\pi$, the probability of an active linear effect, with the proportionality constant depending on what parents are active. Thus (9) becomes

$$P(\delta_{AB} = 1|\delta_A, \delta_B) = \begin{cases} \pi c_0 & \text{if } (\delta_A, \delta_B) = (0,0) \\ \pi c_1 & \text{if one of } \delta_A, \delta_B = 1 \\ \pi c_2 & \text{if } (\delta_A, \delta_B) = (1,1) \end{cases} . \tag{22}$$

with proportionality constants $c_0, c_1, c_2$ and (10) becomes

$$P(\delta_{A_Q} = 1|\delta_{A_L}) = \begin{cases} \pi c_3 & \text{if } \delta_A = 0 \\ \pi c_4 & \text{if } \delta_A = 1 \end{cases} . \tag{23}$$

with proportionality constants $c_3$ and $c_4$. Bingham & Chipman (2002) choose $(c_0, c_1, c_2) = (1.0, 0.5, 0.01)$. For quadratics one might choose $(c_3, c_4) = (0.01, 1.0)$. These choices reduce selection of a prior distribution on $\boldsymbol{\delta}$ to specification of a single hyperparameter $\pi$.

The rationale for these choices is as follows: $c_0 = c_3 = 1.0$ corresponds to the belief that if all parents of an effect are active, then that effect has the same probability of activity as one of its parents. At the other extreme with $c_2 = c_4 = 0.01$, when none of the parents of an effect are active, then it is highly unlikely that the effect will be active. The remaining choice, $c_1 = 0.5$, corresponds to weak heredity, in that an effect with one out of two active parents has some chance of activity, but it should be smaller than if both parents were active.

The hyperparameter $\pi$ may be selected by considering the prior expected number of active effects. Illustrative calculations are given for a full second-order model with $m$ factors, and subsets including linear and quadratic main effects, and linear×linear interactions. This would imply $m$ linear effects, $\binom{m}{2}$ linear×linear interaction effects and $m$ quadratic effects. Prior probability on the subsets would have the form of (22) and (23) above. A straightforward extension of the calculations of Bingham & Chipman (2002) yields an expected number of

| E(# effects) | $\pi$ | E(# linear effects) | E(# linear×linear int.) |
|---:|---|---:|---:|
| 2 | 0.113 | 1.245 | 0.754 |
| 4 | 0.185 | 2.040 | 1.960 |
| 6 | 0.243 | 2.674 | 3.326 |

Table 7: Simulated example: Desired number of effects, corresponding hyperparameter $\pi$ and breakdown by linear effects and two factor interaction effects. There are 11 linear effects and 55 linear×linear interactions.

active effects as

$$\begin{aligned} \text{E}(\#\text{ active effects}) &= m\pi + m\pi\left\{(1-\pi)c_3 + \pi c_4\right\} \\ &\quad +\pi\binom{m}{2}\left\{c_0 + 2\pi(c_1 - c_0) + \pi^2(c_0 - 2c_1 + c_2)\right\} \end{aligned} \tag{24}$$

The two terms on the first line of (24) represent the expected number of linear and quadratic main effects; the second line is the expected number of linear×linear interaction effects.

This simple expression is invaluable for elicitation of a prior distribution: instead of specifying the probability of activity for a variety of different effects, an expected number of effects can be specified, along with values of $c_0 - c_4$ (quite likely the defaults in (22) and (23)). The expression for the expected number of active effects (24) is then solved for $\pi$. A data analyst could also experiment with alternate values of $c_0 - c_4$ and see the impact of these choices in terms of the expected number of linear effects and interactions.

Choices of $\pi$ yielding 2, 4 and 6 expected active effects for the simulated example are shown in Table 7.

# 5 Examples

## 5.1 Simulated data

The data were introduced in Section 4.2. Since all 11 factors ($A$-$K$) are set at two levels, linear main effects and linear×linear interactions are considered. All corresponding contrasts are coded as ±1. The weak heredity prior distribution on $\boldsymbol{\delta}$, (22), is calibrated with $\pi = 0.185$ so that there are 4 expected linear effects. The default choice (20) of $\tau$ and $c$ yields $\tau = 1/3(1-(-1)) = 1/6, c = 10$. For the prior distribution on $\sigma$, default choices are $\lambda = s^2/25 = 10.01/25 = 0.40$ and $\nu = 5$.

1000 draws from the posterior distribution (14) were collected via the Gibbs sampler. The probability that each effect is active is plotted in the top panel of Figure 6, as a vertical line.

This is quite similar to Figure 1 in Section 1.1; details are given below. It is quite clear that the active effects (A, AB, AC) are well identified by the algorithm and all other effects are correctly identified as inactive. Marginal probabilities plotted in Figure 6 are calculated analytically using (19), rather than the relative frequency estimates based on (15).

The joint probability distribution on $\boldsymbol{\delta}$ can also be informative. In this case, the true model $(A, AB, AB)$ dominates, with 50.1% of posterior probability. The two next most probable models each have probability of about 3%. Each involves addition of either $B$ or $C$ linear effects to the most probable model.

An important feature of any subset selection procedure is that when no effects are active, this be identified. To explore this, the 12 values of $\mathbf{Y}$ are randomly shuffled in the order 6,2,7,8,9,1,4,12,5,3,11,10, while rows of $X$ are not changed. The analysis is re-run and the marginal probabilities plotted in the bottom panel of Figure 6. Although a few factors have some probability of activity, there is nothing quite as convincing as in the original analysis.

To explore the sensitivity of the algorithm to a variety of hyperparameter choices, various combinations of $\pi, \tau, c$ were considered. The three values of $\pi$ given in Table 7 were used, giving 2, 4 and 6 expected active effects. Six combinations of $\tau, c$ were explored. The first three choices are $c = 10$ and $\tau = (0.5, 1, 1.5)/6$. These are close to the default hyperparameter choice (20) of $c = 10, \tau = 1/6$. The second three choices are $c = 100$ and $\tau = (0.05, 0.1, 0.15)/6$. These are close to the default hyperparameter choice (21) of $c = 100, \tau = 0.1/6$. The rectangles in Figure 6 represent the range of posterior probabilities over the 18 combinations of hyperparameters. In the top plot, there is minimal sensitivity to the hyperparameters, as the boxes are narrow and most probabilities are near 0 or 1. In the lower plot (with no signal), there is considerably more uncertainty and no effects that are clearly active.

One issue with stochastic simulation methods is how long they should be run. This can be partly addressed by estimating the posterior probability of all models visited so far. To estimate this probability, the normalizing constant $C$ was estimated via the capture-recapture method ((18) in Section 3.2). The "capture set" $\mathcal{A}$ was chosen as the first 1000 draws from a run of 10,000 iterations. The capture set $\mathcal{A}$ contains 364 different values of $\boldsymbol{\delta}$ (that is, 364 distinct subsets of active effects). The other 636 $\boldsymbol{\delta}$ values visited in the first 1000 iterations were duplicates of these 364, and are not included in $\mathcal{A}$. In the remaining 9000 iterations, 73% of the $\boldsymbol{\delta}$ values visited were contained in $A$. Thus in (18), $\sum_{i=1}^{K} I_{\mathcal{A}}(\boldsymbol{\delta}^k)/K = 0.73$. After calculation of the estimated normalizing constant via (18), it is estimated that 82% of the models have been visited by the end of the 10,000 iterations. The estimated cumulative probability of models visited is graphed in Figure 7. The algorithm takes approximately 100 iterations to identify a high-probability value of $\boldsymbol{\delta}$. By 1000 iterations there appear to be very few high probability models that have not been visited, since the slope of the curve has decreased (and continues to

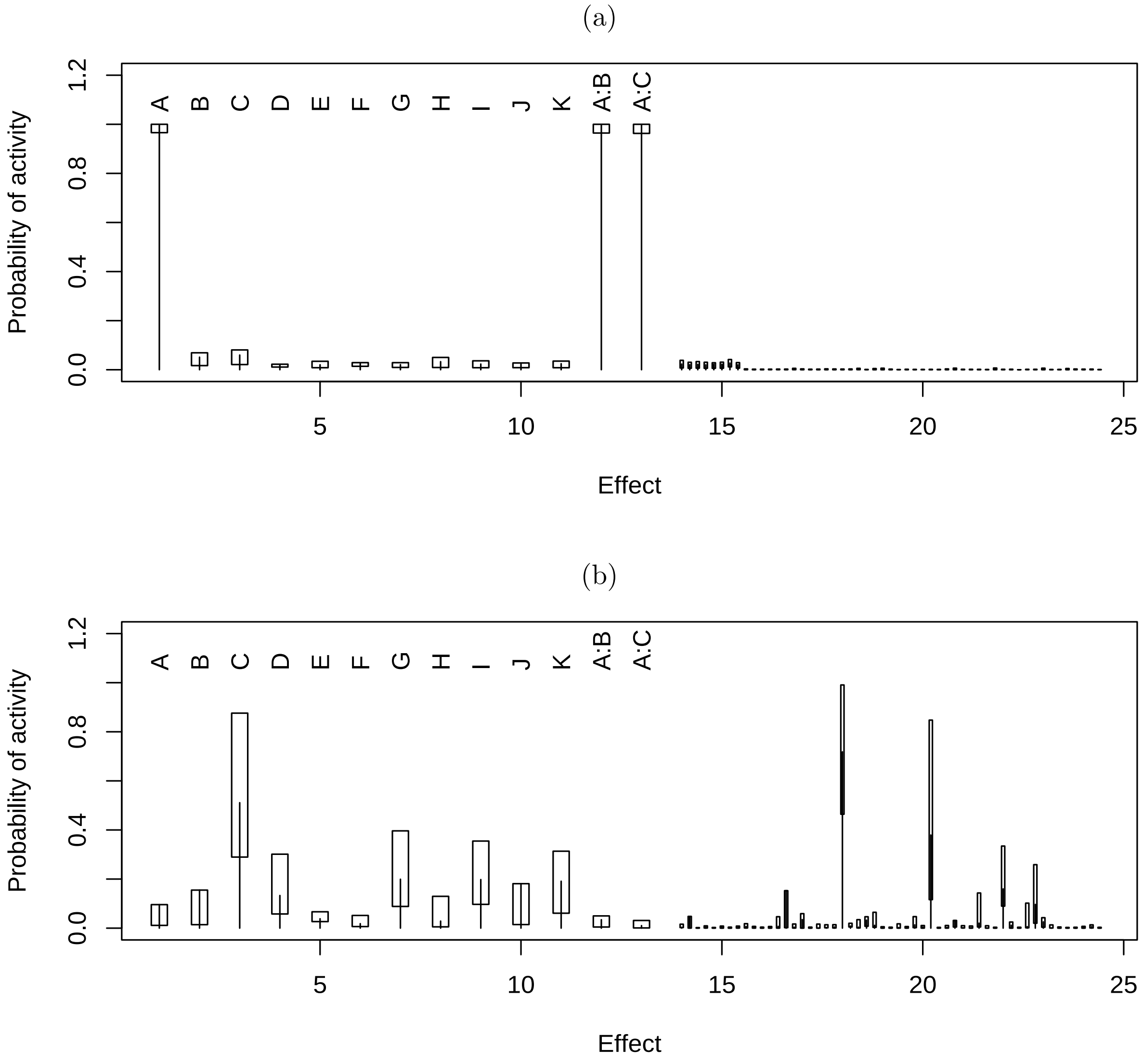

Figure 6: *(a): Marginal probability of activity for each effect, simulated example. Line corresponds to hyperparameter settings giving 4 expected active effects (*$\pi = 0.185, c_0 = 1, c_1 = 0.5, c_2 = 0.01$*), default choice of* $\tau$ *and* $c = 10$*. Boxes represent extremes over 2, 4, 6 prior expected effects and six* $(c, \tau)$ *multipliers of* $(1, .5), (1, 1), (1, 1.5), (10, .05), (10, .1), (10, .15)$*. (b): Same plot, except the response* $\mathbf{Y}$ *has been permuted so there is no signal in the data.*

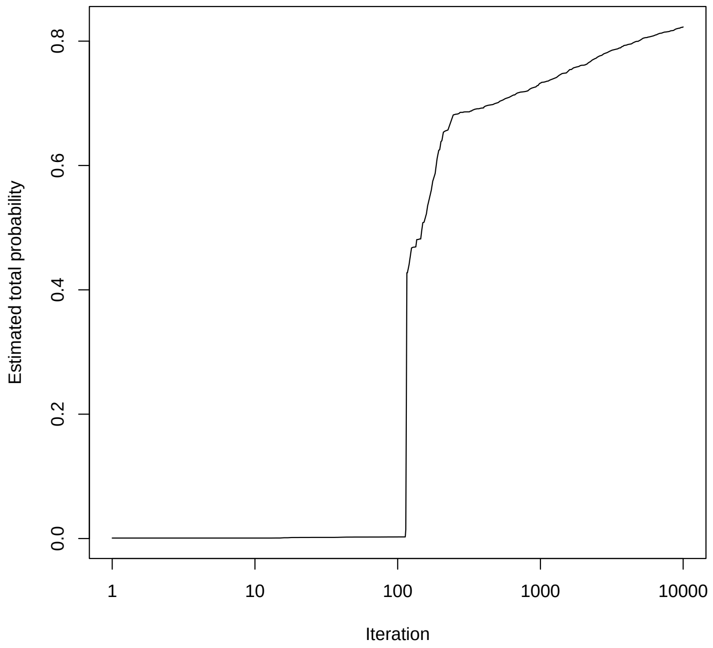

Figure 7: *Simulated data: Estimated cumulative probability of distinct models visited so far.*

decrease, since the horizontal axis is on a $\log_{10}$ scale). There is little advantage of running many more than 1000 iterations for this problem.

Figure 7 also illustrates the "burn-in" problem with relative frequency estimates of posterior probability discussed in Section 3.2. The first 100 iterations of the algorithm visit improbable subsets, so a relative frequency estimate of posterior probability (15) will place too much probability on these 100 subsets.

## 5.2 Glucose data

The main results of the analysis of the glucose data have already been presented in Section 1.1. Details of the hyperparameter choices, robustness calculations, and estimation of the total probability visited by the MCMC sampler are given here.

Prior hyperparameters are set as follows for this example, unless otherwise indicated: $\nu = 5, \lambda = s^2/25 = 101.2282/25 = 4.049$; $c = 100$ and $\tau_j$ are specified according to (21). The exact values of $\tau_j$ will vary with effect index $j$, since the contrasts are coded with different ranges. As discussed in Section 4.2, $c = 10$ seems to allow too much flexibility to for the inactive effects to capture residual error. Calibration of $\pi$ via an expected number of effects is difficult, since effects of so many types (linear, quadratic, linear×linear , linear×quadratic , quadratic×quadratic ) are present. To simplify calculations, calibration is carried out using only the number of linear, quadratic and linear×linear interactions. There are 8 possible linear effects, 7 quadratics and $\binom{8}{2} = 28$ linear by linear interactions, for a total of 43 possible effects. Choosing $\pi = .2786$ gives 5 effects out of the 43 expected to be active. Higher order interactions will raise this expectation, but not much, since all their parents are of at least second order.

A single run of the MCMC sampler is used, with 2500 iterations. The posterior probabilities of models given in Table 4 are normalized so all subsets visited have total probability 1.0. That is, estimate $\widehat{C}$ from (17) is used in conjunction with analytic expression (16) for posterior probability on $\boldsymbol{\delta}$.

In a study of robustness, choices of $\pi = .1486, .2786$ and $.3756$ are considered, giving 2, 5 and 8 expected effects (considering up to linear×linear effects only, as above). Six combinations of $c$ and $\tau$ settings are used, as in the last example, with three $\tau_j$ values at 0.5, 1.0, and 1.5 times the defaults in (20) with $c = 10$ and (21) with $c = 100$. The vertical lines in Figure 1 correspond to the default choices given at the start of this section, and rectangles to ranges over the 18 different hyperparameter settings.

The overwhelming conclusion from both Figure 1 and Table 4 is that that high order interactions between $B$ and $H$ are present. As Chipman et al. (1997) mention, this could well be due to the choice of the original factors: products of volume ($B$) and dilution ($H$) might give some absolute amount of blood in the sample. A transformation might eliminate the need for higher order effects.

A long run of 50,000 steps was carried out to estimate the posterior probability of subsets visited by the MCMC algorithm, via the capture-recapture method (18). The first 500 iterations determined a capture set $\mathcal{A}$, and the remaining 49,500 iterations were used to estimate the total posterior probability visited. Figure 8 shows that just slightly over 40% of the probability is visited by 50,000 steps. This is due to a rather diffuse posterior distribution. The small portion of posterior probability visited by the search at 2,500 iterations (roughly 25%) implies that subset probabilities in Table 4 are likely to be 1/4 the values given. What is perhaps more important is the relative size of the probabilities, given that the posterior distribution is quite diffuse. In this problem more information may be obtained from marginal posterior distributions on individual $\delta_j$ than the joint posterior distribution on $\boldsymbol{\delta}$.

# 6 Prior distributions for design

Bayesian methods have often proven useful for design of experiments, especially in situations in which the optimal design depends on unknown quantities. Certainly, to identify a design for optimal estimation of $\boldsymbol{\beta}$, the correct subset of effects must be identified. Bayesian approaches

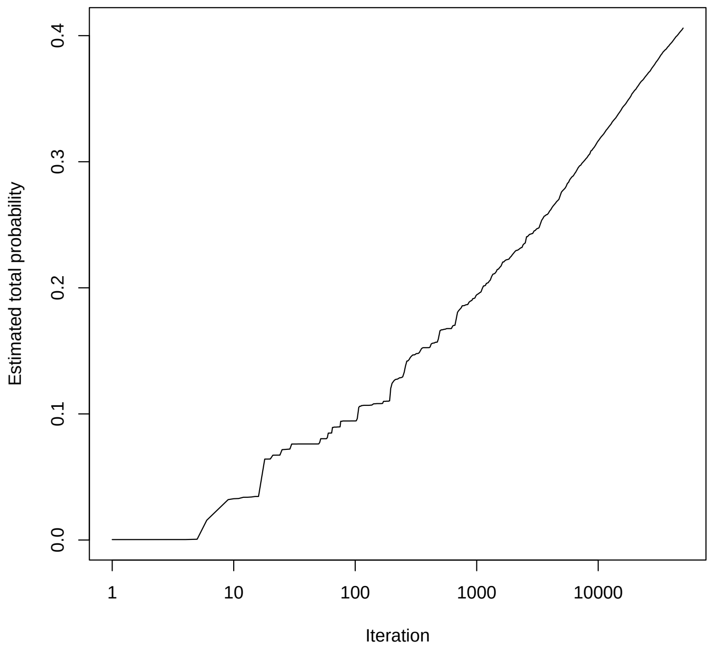

Figure 8: *Glucose Data: Estimated cumulative probability of distinct models visited, up to 50,000 iterations.*

that express uncertainty about the correct subset enable construction of optimality criteria that account for this uncertainty, typically finding a design that optimizes a criterion which is averaged over many possible subsets. DuMouchel & Jones (1994) exploit this idea with a formulation in which some effects have uncertainty associated with whether or not they are active. Meyer, Steinberg & Box (1996) extend the prior distributions of Box & Meyer (1993) and construct a "model discrimination" design criterion. The criterion is based on a Kullback-Leibler measure of dissimilarity between predictions from two competing models and averages this dissimilarity over all possible pairs of models. Averaging is weighted according to the prior probability of the models, thus incorporating prior information into the design criterion. Bingham & Chipman (2002) use weak heredity (22) in a similar criterion based on the Hellinger distance.

Even in seemingly straightforward cases, such as a 16-run design, prior information can lead to non-regular designs. For example, Bingham & Chipman (2002) found that if sufficiently small values of $\pi$ were used when looking for a 6-factor, 2-level design in 16 runs, a non-regular fractional factorial design was optimal. Non-regular designs were chosen over regular fractional factorials because they are better at estimating models containing a mix of main effects and interactions. Regular fractional factorials enable estimation of many linear effects, at the cost of estimability of interactions. Some large linear-effect-only models may actually seem implausible. For example, prior distribution (22) with $\pi = 0.41$ puts over 600 times more prior probability on a model with effects $A, B, AC$ than a model with linear effects $A, B, C, D, E, F$. This leads the design criterion to select designs that sacrifice simultaneous estimability of all linear effects for the ability to estimate more interactions.

# 7 Discussion

Although the presentation here has focused on linear models with Gaussian errors, similar ideas may be applied to subset selection in other models. For example, George, McCulloch & Tsay (1995) extend the nonconjugate $\boldsymbol{\beta}$ prior distribution (5) to a probit regression model for a binary response. They exploited a relationship between a probit regression and a linear regression with normal errors. Let $Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon$, with $\epsilon \sim N(0, 1)$. Instead of observing $Y_i$ we observe a binary response $Z_i$ that is a thresholding of $Y_i$. Thus, we observe $Z_i = 1$ if $Y_i > 0$ and $Z_i = 0$ otherwise. Then $\Pr(Z_i = 1) = \Pr(Y_i > 0) = \Phi(\mathbf{x}_i'\boldsymbol{\beta})$ with $\Phi$ being the standard normal cumulative probability function. This is the probit model for a binary response. George et al. (1995) treated $Y$ as missing data and use the data augmentation approach of Tanner & Wong (1987) to simulate the unobserved $Y$. The Gibbs sampler for (5) can be applied to $Y$, so the extended algorithm alternates between draws of $\boldsymbol{\beta}$ and the unobserved latent variable $Z$.

Other more general modifications and extensions are possible. If a MCMC sampler for a model exists, then an additional step incorporating a draw of the subset indicator $\boldsymbol{\delta}$ should be possible. For example in a generalized linear model with regression coefficient vector $\boldsymbol{\beta}$, dispersion parameter $\phi$ and a fixed subset of active effects, a MCMC sampler might be available for the full joint posterior distribution $p(\boldsymbol{\beta}, \phi|\mathbf{Y})$. To generalize to subset selection, it will be necessary to draw from $p(\boldsymbol{\beta}, \phi, \boldsymbol{\delta}|\mathbf{Y})$. The draws for $\phi$ will be unchanged. The draws for $\boldsymbol{\beta}$ will be the same, except the prior variances will be determined by $\boldsymbol{\delta}$. The draw for $\boldsymbol{\delta}$ will be carried out one element at a time using the conditional distribution $p(\delta_j|\delta_1, \ldots, \delta_{j-1}, \delta_{j+1}, \delta_p, \boldsymbol{\beta}, \phi, \mathbf{Y})$. This conditional probability distribution will be a binary draw, with the probability of $\delta_j = 1$ depending on the ratio of two densities of $\beta_j$ (one with $\delta_j = 0$, the other with $\delta_j = 1$). One paper that develops such a sampler (without subset selection) is Dellaportas & Smith (1993), in which a Gibbs sampling scheme is used for generalized linear models and proportional hazards models.

Analytic approaches, in which the marginal posterior distribution of $\boldsymbol{\delta}$ is obtained by integrating the posterior distribution with respect to $\boldsymbol{\beta}, \phi$, are also possible. Analytic approximations such as the Laplace approximation (Tierney & Kadane 1986) would be necessary to obtain a closed form expression for the marginal posterior distribution $p(\boldsymbol{\delta}|\mathbf{Y})$.

Another interesting problem in which Bayesian subset selection prior distributions might be used is in situations in which there is complete aliasing between effects. By adding information about relationships between effects in the form of heredity prior distributions, the posterior distribution could be used to disentangle the most likely effects. Chipman & Hamada (1996) discover such a pattern, in which there is support for the model $A, C, E, H, AE = CH = BF = DG$. The last four effects are aliased. Two of these $(BF, DG)$ are discarded automatically because they do not obey heredity. Two sub-models involving the other two are identified: $C, E, H, CH$ and $A, C, E, AE$, with the former providing better fit and consequently receiving higher posterior probability.

The emphasis of this paper is very much on subset *selection*, since the goal of screening experiments is to identify factors that influence the response. Bayesian model averaging is an alternative, in which predictions are averaged across all possible subsets (or a representative sample), using posterior probability as weights. The consensus among researchers in the field seems to be that model averaging produces better accuracy than selection of a single subset. However, in screening experiments, selection remains paramount. A review of Bayesian model averaging is given in Hoeting, Madigan, Raftery & Volinsky (1999).

# References

Bingham, D. & Chipman, H. (2002), Optimal designs for model selection, Technical Report 388, University of Michigan.

Box, G. E. P. & Meyer, R. D. (1986), ‘An analysis for unreplicated fractional factorials’, *Technometrics* **28**, 11–18.

Box, G. E. P. & Meyer, R. D. (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’, *The Canadian Journal of Statistics* **24**, 17–36.

Chipman, H. A. (1998), Fast model search for designed experiments, *in* B. Abraham, ed., ‘Quality Improvement Through Statistical Methods’, Birkhauser, Boston, MA, pp. 205–220.

Chipman, H. A., George, E. I. & McCulloch, R. E. (2001), The practical implementation of Bayesian model selection, *in* P. Lahiri, ed., ‘Model Selection’, Vol. 38 of *IMS Lecture Notes - Monograph Series*, Institute of Mathematical Statistics, Beachwood, OH, pp. 65–116.

Chipman, H. A. & Hamada, M. S. (1996), ‘Comment on “Follow-up designs to resolve confounding in multifactor experiments”’, *Technometrics* **38**, 317–320.

Chipman, H., Hamada, M. & Wu, C. F. J. (1997), ‘A Bayesian variable-selection approach for analyzing designed experiments with complex aliasing’, *Technometrics* **39**, 372–381.

Dellaportas, P. & Smith, A. F. M. (1993), ‘Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling’, *Applied Statistics* **42**, 443–459.

Draper, N. R. & Smith, H. (1998), *Applied regression analysis*, John Wiley & Sons.

DuMouchel, W. & Jones, B. (1994), ‘A simple Bayesian modification of *D*-optimal designs to reduce dependence on an assumed model’, *Technometrics* **36**, 37–47.

Furnival, G. M. & Wilson, Robert W., J. (1974), ‘Regression by leaps and bounds’, *Technometrics* **16**, 499–511.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (1995), *Bayesian data analysis*, Chapman & Hall Ltd.

George, E. I. & McCulloch, R. E. (1993), ‘Variable selection via Gibbs sampling’, *Journal of the American Statistical Association* **88**, 881–889.

George, E. I. & McCulloch, R. E. (1997), ‘Approaches for Bayesian variable selection’, *Statistica Sinica* **7**, 339–374.

George, E. I., McCulloch, R. E. & Tsay, R. S. (1995), Two approaches to bayesian model selection with applications, *in* D. A. Berry, K.A.Chaloner & J. K. Geweke., eds, ‘Bayesian

Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner', John Wiley, New York, NY, pp. 339–348.

Hamada, M. & Wu, C. F. J. (1992), 'Analysis of designed experiments with complex aliasing', *Journal of Quality Technology* **24**, 130–137.

Henkin, E. (1986), The reduction of variability of blood glucose levels, *in* 'Fourth Supplier Symposium on Taguchi Methods', American Supplier Institute, Dearborn, MI, pp. 758–785.

Hoeting, J. A., Madigan, D., Raftery, A. E. & Volinsky, C. T. (1999), 'Bayesian model averaging: a tutorial (Disc: 402–417, Corr: 3V15 p193–195)', *Statistical Science* **14**(4), 382–401.

Lee, P. M. (1997), *Bayesian statistics: an introduction*, Edward Arnold Publishers Ltd.

Meyer, R. D., Steinberg, D. M. & Box, G. (1996), 'Follow-up designs to resolve confounding in multifactor experiments (Disc: p314-332)', *Technometrics* **38**, 303–313.

Nelder, J. A. (1998), 'The selection of terms in response-surface models — How strong is the weak-heredity principle?', *The American Statistician* **52**, 315–318.

Peixoto, J. L. (1990), 'A property of well-formulated polynomial regression models (Corr: 91V45 p82)', *The American Statistician* **44**, 26–30.

R Development Core Team (2004), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3.

Raftery, A. E., Madigan, D. & Hoeting, J. A. (1997), 'Bayesian model averaging for linear regression models', *Journal of the American Statistical Association* **92**, 179–191.

Tanner, M. & Wong, W. H. (1987), 'The calculation of posterior distributions by data augmentation', *Journal of the American Statistical Association* **82**, 528 – 550.

Tierney, L. & Kadane, J. B. (1986), 'Accurate approximations for posterior moments and marginal densities', *Journal of the American Statistical Association* **81**, 82–86.

Zellner, A. (1987), *An introduction to Bayesian inference in econometrics*, Krieger Publishing Company, Inc.