Bayesian Regression Methods for Ordered Categorical Data

by

Hugh A. Chipman

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 1994

I hereby declare that I am the sole author of this thesis.

I authorize the University of Waterloo to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the University of Waterloo to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Abstract

This thesis explores regression models for ordered categorical data, with special attention to data arising from industrial experiments. There has been considerable interest in models to deal with this kind of response, and significant advances have been made by Nair (1986) and Hamada and Wu (1990). These papers provide an in-depth look at the existing methods for analyzing ordinal data, including accumulation analysis and scoring, clarifying their strengths and weaknesses. Despite the advances made, both papers still rely on assigning a numeric score to each response category, and analyzing the scores as though they were normal. This heuristic approach is an approximation, and in some cases can be less than optimal. One alternative suggested by Hamada and Wu is to fit models to the data via maximum likelihood instead. This is a more sound approach, and the regression models of McCullagh (1980) for ordinal data would seem appropriate, since they are an extension of binary regression.

In the industrial setting, where models for experimental data typically have as many terms as observations, the MLE's quite often do not exist. This renders the maximum likelihood approach to fitting McCullagh models inappropriate when the models are to be used for optimization. An exciting alternative procedure is to assume prior knowledge about the effects of the variables, and fit the model via Bayesian techniques. This is reasonable, since the estimated effects are known to be finite on the transformed scale. The fitting of Bayesian McCullagh models is itself a challenge, since the posterior distributions are not available in closed form. The Gibbs sampler, a recent computational technique, is used to obtain the posterior distributions of the parameters and any functions of them. A second important step is considered in addition to modeling, namely optimization of the process using the fitted model. In addition to optimization of plain factorial experiments, a response model strategy (Welch, Yu, Kang and Sacks, 1990 and Shoemaker, Tsui and Wu, 1991) for optimization of Taguchi's "robust parameter design" experiments is developed. An important advantage of the Bayesian approach is that the posteriors capture model uncertainty, which is incorporated in the optimization process to produce a more realistic choice of optimum levels.

One of the key issues in fitting this model is variable selection. Unlike ANOVA methods, which can determine variable importance marginally, it is necessary to search the entire model space. A generalization of the basic model that incorporates automatic Bayesian variable selection (George and McCulloch, 1993) is explored. This method searches the entire model space, and identifies the most probable models, with the a priori assumption that selection of terms is independent. In industrial problems, many terms in the model are related to one another, so that independence is not a reasonable assumption. For example, if neither main effect A or B is identified as important, models containing an interaction AB may be considered implausible. Prior structures that represent this and other relations among predictors are developed and applied to industrial problems and data. The relation considered include dummy variables, polynomial predictors, mutually exclusive predictors, and restrictions on the numbers of active predictors. Although the variable selection part

of the thesis is developed for ordered categorical data, it may be applied to other regression problems, including linear models.

Acknowledgements

I would like to thank Professor Jeff Wu and Professor Mike Hamada for their guidance, advice, and invaluable assistance during my graduate studies at Waterloo.

I thank Professors Greg Bennett, Jerry Lawless, Robert McCulloch, and Park Reilly for serving on my thesis committee.

I thank Greg Anglin, Professor Bruce Hill, Julie Horrocks, Michael Lewis, Professor Vijay Nair, and Professor Martin Young for stimulating conversations related to this research.

I thank the University of Waterloo and the University of Michigan for making their research and computing facilities available to me.

I thank the Natural Sciences and Engineering Research Council of Canada and the Manufacturing Research Corporation of Ontario, who in part supported this work.

I thank my wife, Tania for her support and encouragement.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Data sets with an ordered categorical response arise naturally in a number of industrial situations. Such a response can be characterized as a random variable that takes on a number of discrete outcomes or categories, which have an implicit ordering, such as "poor", "average", and "good". While there may be any number of categories, a small number (two to six) is common. Binary responses are an important special case.

In industrial situations, such data may be quick surrogates for continuous measurements. This occurs when the main priority is to run the experiment quickly, and if necessary consider a more complete measurement later. For example, engineers might be interested in the force required to close a door. While devices exist to make such measurements, they may be expensive or time-consuming, so an expert might be used to make categorical judgments, such as "difficult", "acceptable", and "easy". In other situations, such as in assessing the physical appearance of steel, the appearance can be graded but there is no obvious underlying continuous response. Moreover, it might require much ingenuity and cost to develop a device which could measure the appearance appropriately on some continuous scale.

The data considered in this thesis have several features specific to industrial experiments. For example, the predictor variables (or factors) are typically set at a relatively small number of levels, and are arranged in an orthogonal array. Also, the number of predictor terms is typically close to the number of observations made. This means that relatively simple models are used, and identification of important terms will be a priority. Although industrial data are considered, most of the methods developed will be applicable to more general designs and problems.

Chapter 2 looks at models to describe the dependence of the ordered categorical response on predictor variables. Widely available software allows the fitting of several models, including Taguchi's (1974) accumulation analysis (AA), scored ANOVA (Nair, 1986), and generalized linear models (GLMs) (McCullagh 1980, McCullagh and Nelder 1989). AA has serious problems (as mentioned in Nair (1986) and Hamada and Wu (1990)), and is not recommended. Although an improvement over AA, scored ANOVA relies on assigning a number to each outcome category and analyzing this pseudo response as if it had normal errors. The sums of squares tend to underestimate the variance of the response, which can cause spurious effects to be identified as significant. GLMs are a more promising alternative, because they model the outcome counts instead of a pseudo response. The routine application of GLMs is limited for factorial experiments, because the sparsity of data will often produce MLE's which are infinite or fail to converge. Not only is current software ill-equipped to handle this problem, but optimization of the process using a model with infinite MLE's can be difficult or impossible. One area where optimization is not possible is the response model approach for Taguchi's robust design experiments.

Although the method of maximum likelihood appears inappropriate for fitting GLMs to ordinal data, the GLM model itself is a good idea. In this thesis, a Bayesian method for fitting the GLM with informative priors is proposed to circumvent the problem of infinite MLE's, while at the same time retaining the superior modeling ability of GLMs. Previously, Hamada and Wu (1992) showed that similar benefits are realized by using a Bayesian approach for analyzing censored data. One major challenge in fitting a Bayesian GLM is the numeric calculation of posteriors, which are not available in closed form. A numeric technique, the Gibbs sampler (Geman and Geman (1984) and Gelfand and Smith (1990)), is used to obtain samples from the posteriors which may be used for inference. The details are similar to Albert and Chib (1993), who consider Bayesian models for various types of categorical data. The calculations necessary to implement this model are given in Section 2.2.2, after a brief overview of Bayesian computational methods in Section 2.2.1.

Chapter 2 also indicates how these models will be fit and used, in Section 2.3. Variable selection is the main challenge in fitting these models, and an informal, non-automated strategy is outlined in Section 2.3.1. An important component of variable selection is a criterion for assessing the fit of the model; several are reviewed in Section 2.3.2. Once the model is fit, it can then be used to optimize the process. Optimization consists of finding settings for the predictors which give high probability of a response in the desired category. An important by-product of using the Bayesian approach is that the posteriors capture model uncertainty. The optimization is over these posteriors and hence produces more realistic optimal values. An additional optimization problem is considered, namely Taguchi's robust design experiments. Here, some of the predictors are uncontrollable under regular production, and the goal is to choose levels of the other controllable predictors to desensitize the response to the uncontrollable predictors. The variation in the uncontrollable predictors is easily combined with the posterior samples in the optimization process.

The fitting of this model and subsequent optimization are illustrated in Chapter 3. Two industrial data sets are considered, one a robust design with a three level response, and the other a simple factorial experiment with a six level response. Several specific techniques and graphics not developed in Chapter 2 are considered here, including practical convergence diagnostics, graphics to assist optimization, robustness of the conclusions to the prior, and on-target optimization. One shortcoming of the proposed method becomes evident in these examples, namely variable selection.

The second half (Chapters 4 to 7) of the thesis develops an alternative variable selection method which offers substantial advantages over the previously suggested approach.

These advantages include flexibility in indicating preference towards certain models, a more accurate representation of relations between terms, and the ability to rapidly search the entire model space. The automated Bayesian variable selection algorithm of George and McCulloch (1993) forms the basis for this approach. The basic idea is to add binary-valued parameters to the model to indicate whether each individual term is "active" or "inactive". The joint posterior for these parameters will then indicate which models are most probable given the data. Although their theory is general, in practice George and McCulloch use an independence prior on these parameters, assuming that a term is active (or inactive) independently of all others. This is not the case in industrial experiments, where knowledge that certain terms are active may alter the probability that other related terms are active. For example, if terms A and B appear active, then an AB interaction seems more likely to be active than if neither A or B were active. This idea forms the basis of the generalization of George and McCulloch's method to related predictors. Several different relations between terms are possible, and these are explored in Chapter 5. Some of these relations are quite common, and will be referred to as "principles" to emphasize the fact that the priors are a mathematical representation of beliefs that are commonly held in variable selection. Priors which represent these structures are developed in each case, including two-way interactions, general polynomial interactions, grouped terms, competing predictors, and constraints on the number of terms allowed in the model. Several of these priors are incorporated into the software developed in the first part of the thesis; pertinent computational details are provided in the last section of this chapter.

The applicability of the variable selection technique to several industrial problems is illustrated in Chapter 6. With robust design experiments, there is a special interest in interactions between controllable and uncontrollable variables, as they can be used to in optimization to reduce variability. In the case of non-interacting subsystems, priors incorporate the information that interactions between variables in different systems appear unlikely. Priors that limit the number of terms in a model find application to oversaturated designs, which have more predictors than observations. Finally, in situations where there is almost complete aliasing, competing predictor priors can be used to allow only term in a group of highly correlated terms to be active.

The applications discussed in Chapter 6 are not actually applied to real data. To illustrate the strengths of this method, a data set is examined in Chapter 7. The design of the experiment allows illustration of several of the principles of Chapter 5. Both the priors and the posteriors are examined, the former as a "reality check" to ensure that priors are reasonable, and the latter to determine which models seem probable given the data. Several new graphics are introduced to facilitate the examination of the high dimensional distributions.

A summary of the important findings of this thesis is presented in Chapter 8.

Chapter 2

Models and their Application

2.1 Current Models

Several techniques are available for the analysis of ordered categorical data. Among those used for industrial situations are scored ANOVA, accumulation analysis, and generalized linear models. This section outlines these techniques and some of the problems they have.

The most straightforward method of analysis assigns arbitrary (strictly increasing) scores to the ordered response categories, and performs an ANOVA on these scores. While this approach is familiar, there are some complications. Any conclusions made will depend on the scores used as discussed in Hamada and Wu (1990). Furthermore, the scored data are not continuous and the normal distribution may be an unreasonable distributional approximation when there are few categories (say less than six). This poor approximation is manifested in the sums of squares, which tend to underestimate the residual variance because they are based on a discrete response. This leads to inflated significance levels, and the possibility of identifying unimportant terms as significant. A concrete example of this is given in a case study in Section 3.1.

Another technique, Accumulation Analysis (AA), is an adapted ANOVA which uses the cumulative frequencies of the ordered categories as a response. It was introduced by Taguchi (1974) as an improvement over Pearson's chi-squared test, which does not use the order of the categories. There are a number of serious criticisms raised against AA by Nair (1986), Hamada and Wu (1990) and discussants therein, which include detection of spurious effects, testing for a combination of location and dispersion effects, and reversing the order of factor importance. The method's shortcomings make further discussion unnecessary, and the reader is referred to the papers for further details.

Many of the problems with ANOVA based techniques relate to the validity of the inference they make. A search for statistically valid methods leads naturally to Generalized Linear Models (GLMs) for ordinal data. These models are attractive since software is readily available for fitting them.

The model may be summarized as follows: instead of creating a pseudo measurement to be analyzed, the probabilities of the J ordered categories are modeled as a function of the predictors. McCullagh (1980) suggests a family of models of the form:

$$\Pr(Y_i \leq j) = \Phi(\gamma_j - X_i'\boldsymbol{\beta}) \qquad \text{for} \qquad j = 1, ..., J-1 \tag{2.1}$$

where categories are labeled $1, \ldots, J$, the inverse link function Φ is a (known) monotone increasing function mapping the real line $(-\infty, \infty)$ onto the interval (0,1), γ_j is a "cutpoint", X_i is a vector of covariates including main effects and interactions, and $\boldsymbol{\beta}$ is a vector of effects. The γ_js may be thought of as intercepts, since there is no intercept term in the vector $\boldsymbol{\beta}$. A natural ordering of cells is obtained by modeling cumulative cell probabilities rather than individual probabilities. Although each response category has a corresponding cutpoint, the regression coefficients $\boldsymbol{\beta}$ are constant across outcomes.

Once a link function is specified, estimation of the parameters $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ is carried out via maximum likelihood (ML). One problem with this approach is that estimates for certain coefficients or cutpoints may not exist. For example, suppose we have a binary response and a factor A with two levels. If all of the observations at the low level of A are in the first category, and all the observations at the high level of A are in the second category, then the estimated cell probabilities are zero and one. This corresponds to an estimate of $+\infty$ for the coefficient of A. In multifactor experiments, this problem is quite likely to occur, especially when the number of effects in the model is near the number of runs. These problems will manifest themselves as estimates which fail to converge. Tse (1986) gives conditions under which MLE's are infinite.

Despite the non-convergence of estimates, the likelihood does converge, and may be used as a criterion for assessing the relative importance of the factors. Such an approach is discussed in Lawless and Singhal (1978), and is used in Section 2.3.2. However, in situations where the model is to be used for prediction, or comparisons between effects are necessary, non-existent estimates are a problem.

One solution would be to assume some sort of prior knowledge of the coefficients, and use Bayesian techniques to fit model (2.1) to the data. Even when knowledge about the coefficients is minimal or nonexistent, this approach has justification; it seems plausible that the parameters are only large, and there is not sufficient data to distinguish between large and infinite values. On the probability scale, this prior corresponds to the belief that no response category will have probabilities of exactly zero or one.

Thus, the Bayesian approach retains the advantages of GLMs, which do not require a choice of scores and provide a more accurate description of the response distribution. At the same time, the issue of infinite estimates is resolved; moreover the Bayesian approach is well suited for small sample cases, where the asymptotic normality of ML theory may be a poor approximation. The details of this method are provided in the next section, including the use of Gibbs sampling techniques for implementing the analysis.

2.2 Bayesian Methods

In this section, the theory and tools for fitting a Bayesian GLM are outlined. The main focus is the Gibbs sampler, a computational technique used in the calculation of posteriors and their summaries, such as marginals, moments, or distributions of transformations of the parameters. A brief overview of Bayesian computational methods and the Gibbs algorithm is given before specific results are derived.

Note that here and in the rest of the thesis, the following generic notation will be used: The generic function f will be assumed to represent the density of its argument is, so

$$f(x) \text{ is the density of } x,$$

$$f(y) \text{ is the density of } y,$$

but in general

$$f(x) \neq f(y).$$

2.2.1 A Brief Overview of Bayesian Computational Methods

This section borrows heavily from Smith (1991), which presents a summary of Bayesian computational methods, and Smith and Roberts (1993), which reviews Markov chain Monte Carlo (or MCMC) methods, including the Gibbs sampler. Other important references include Gelfand and Smith (1990), which kindled much of the current interest in the use of MCMC methods, Tierney (1994), which examines several MCMC methods and discusses ways in which they can be combined into "hybrid" methods, and Tanner and Wong (1987), which introduces the idea of data augmentation.

On the surface, Bayesian inference seems quite straightforward. A probability model is specified for the data (e.g. the likelihood $f(\mathbf{Y}|\boldsymbol{\theta})$) and a prior $f(\boldsymbol{\theta})$ for parameters $\boldsymbol{\theta}$. The posterior distribution is simply

$$f(\boldsymbol{\theta}|\mathbf{Y}) = f(\mathbf{Y}|\boldsymbol{\theta})f(\boldsymbol{\theta}) \Big/ \int f(\mathbf{Y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}\ , \tag{2.2}$$

a normalized form of the product of the likelihood and the prior. By integration, information about the posterior, its marginals, and any functionals of it may be obtained. The problem is that the integration can be quite difficult. In general integrals of the form

$$S_I[q(\boldsymbol{\theta})] = \int q(\boldsymbol{\theta})f(\mathbf{Y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}_{I'} \tag{2.3}$$

are required, were I is some index set of the parameter vector. The integration is over the complement of this set I. For example the normalizing constant in (2.2) is obtained by taking

$$I = \emptyset \qquad \text{and} \qquad q(\boldsymbol{\theta}) = 1.$$

The marginal posterior of $\boldsymbol{\theta}_1$ is given by

$$\frac{S_I[1]}{S_\emptyset[1]} = \frac{\int f(\mathbf{Y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\theta_2 \dots d\theta_k}{\int f(\mathbf{Y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}},$$

with

$$I = \{1\}.$$

Similarly, the posterior expectation of some function $q(\boldsymbol{\theta})$ is given by $S_\emptyset[q(\boldsymbol{\theta})]/S_\emptyset[1]$.

A variety of computational methods have been developed for evaluation of integrals such as (2.3). Smith (1991) provides a good overview of commonly used approaches, grouped in the general classes of approximation methods, numeric integration, and Monte Carlo techniques.

Approximation methods attempt to find analytic approximations to (2.2) which are simple to manipulate. The simplest approximation is to assume that for large samples and relatively uninformative priors the posterior (2.2) is approximately $N(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Sigma}})$, where $\widehat{\boldsymbol{\theta}}$ is the MLE and $\widehat{\boldsymbol{\Sigma}}$ the inverse of the second derivative matrix evaluated at $\widehat{\boldsymbol{\theta}}$ respectively. Tierney and Kadane (1986) propose the *Laplace approximation*, a more accurate analytic approximation which uses an expansion about the posterior mode rather than the MLE, and thus incorporates prior information. Smith (1991) points out that this method will be less tractable for high dimensional problems or ones with complex structures. For more complex problems, numeric approximations or Monte Carlo methods are required.

Numerical integration, the second class of approaches, consist of quadrature methods which assume that the integrand of (2.3) can be well approximated by a product of some function and a (known) normal density. Quadrature methods construct an approximation of (2.3) based on evaluation of its integrand at regularly spaced points in the parameter space. Smith (1991) comments that beyond six to nine dimensions, or for complex problems, these approaches are generally not tractable.

Beyond six to nine dimensions, Monte Carlo methods are generally necessary. The basic idea comes from noting that the integral

$$\int q(\boldsymbol{\theta}) h(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{2.4}$$

is the expectation of $q(\boldsymbol{\theta})$ with respect to some generic density $h(\boldsymbol{\theta})$. This suggests approximation of (2.4) with the "sample equivalent", namely

$$\sum_{i=1}^{N} q(\boldsymbol{\theta}^i)/N$$

where

$$\{\boldsymbol{\theta}^i\}_{i=1}^{N} \sim h(\boldsymbol{\theta}).$$

Provided that samples from $h(\boldsymbol{\theta})$ are available, normalized integrals of the form (2.3) may be evaluated, with the arbitrary function $h(\boldsymbol{\theta})$ replaced by the posterior $f(\boldsymbol{\theta}|\mathbf{Y})$.

It can be difficult to sample directly from the posterior, but it may be possible to sample from some other function which closely resembles the posterior. *Importance sampling* methods are a form of Monte Carlo integration which arise from the observation that (2.4) can be written as

$$\int q(\boldsymbol{\theta})h(\boldsymbol{\theta})\frac{g(\boldsymbol{\theta})}{g(\boldsymbol{\theta})}d\boldsymbol{\theta} = \int \left[q(\boldsymbol{\theta})\frac{h(\boldsymbol{\theta})}{g(\boldsymbol{\theta})}\right] g(\boldsymbol{\theta})d\boldsymbol{\theta},$$

for some density $g(\boldsymbol{\theta})$. This suggests the approximation

$$\sum_{i=1}^{N} q(\boldsymbol{\theta}^i)w_i/N \tag{2.5}$$

where

$$\{\boldsymbol{\theta}^i\}_{i=1}^N \sim g(\boldsymbol{\theta}),$$

with "importance weights"

$$w_i = \frac{h(\boldsymbol{\theta}^i)}{g(\boldsymbol{\theta}^i)}.$$

More generally, if h is non-negative, but does not integrate to one (as is the case for the un-normalized posterior $f(\mathbf{Y}|\boldsymbol{\theta})f(\boldsymbol{\theta})$), then the integral (2.4) may be written as

$$\int q(\boldsymbol{\theta})\frac{h(\boldsymbol{\theta})}{\int h(\boldsymbol{\theta})d\boldsymbol{\theta}}d\boldsymbol{\theta} = \int q(\boldsymbol{\theta})\frac{h(\boldsymbol{\theta})}{g(\boldsymbol{\theta})}g(\boldsymbol{\theta})d\boldsymbol{\theta} \Big/ \int \frac{h(\boldsymbol{\theta})}{g(\boldsymbol{\theta})}g(\boldsymbol{\theta})d\boldsymbol{\theta},$$

and approximated by

$$\sum_{i=1}^{N} q(\boldsymbol{\theta}^i)w_i \Big/ \sum_{i=1}^{N} w_i \ .$$

For importance sampling to be effective, the density g must be carefully chosen so that

$$h(\boldsymbol{\theta}) \approx g(\boldsymbol{\theta}).$$

The variance of the estimate (2.5) depends crucially on the closeness of h and g. Thus importance sampling is a weighted version of Monte Carlo integration, and all expectations (like marginals or moments) are calculated using the weights w_i. For an example of the application of importance sampling to a binary data problem, and comparison with quadrature methods, see Zellner and Rossi (1984).

The problem with importance sampling methods is the difficulty of constructing a g which closely matches the posterior distribution. However, when this procedure can be carried out, it is generally more efficient than the dependent methods described below.

When it is not convenient to use importance sampling, an alternative is the class of *Markov chain Monte Carlo* (MCMC) methods. They construct a Markov chain $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots$ with the parameter space as the state space, and limiting distribution f equal to the posterior (or any other distribution from which samples are desired.) They are called dependent sampling methods because realizations $\boldsymbol{\theta}^i$ and $\boldsymbol{\theta}^j$ from the same sequence are correlated.

Simulation of this chain is often quite simple, making this class of techniques attractive. Furthermore, once approximate convergence has occurred, the samples generated are unweighted, simplifying some calculations and making graphical techniques generally more applicable, as noted in Tierney (1994).

Suppose that $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^t, \ldots$ is a realization from some such Markov Chain. Then under weak conditions (typically aperiodicity and irreducibility) typically available asymptotic results include

$$\boldsymbol{\theta}^t \xrightarrow{d} \boldsymbol{\theta} \sim f(\boldsymbol{\theta}); \qquad \frac{1}{t}\sum_{i=1}^{t} q(\boldsymbol{\theta}^i) \rightarrow E_f(q(\boldsymbol{\theta})) \qquad \text{almost surely,} \tag{2.6}$$

as $t \rightarrow \infty$. See Smith and Roberts (1993) for general discussion, and Tierney (1994) or Schervish and Carlin (1992) for more details. As the second property in (2.6) indicates, the samples generated may be used for Monte Carlo integration of functions over the posterior.

There are numerous ways of constructing Markov chains for the purpose of Monte Carlo integration. Tierney (1994) provides a good overview, and suggests ways in which different approaches may be combined to produce hybrid algorithms. For the purpose of this thesis, a specific form of MCMC is used, namely *Gibbs sampling* (Geman and Geman, 1984).

The algorithm may be summarized as follows. Suppose $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ has joint density $f(\boldsymbol{\theta})$, and that

$$f(\theta_i | \boldsymbol{\theta}_{(-i)}) \equiv f(\theta_i | \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k)$$

is the conditional density of θ_i given all the remaining elements of $\boldsymbol{\theta}$. First, arbitrary starting values $\boldsymbol{\theta}^0$ are chosen, and then the following random draws are made:

$$\begin{aligned}
\theta_1^1 &\sim f(\theta_1 | \boldsymbol{\theta}^0_{(-1)}) \\
\theta_2^1 &\sim f(\theta_2 | \theta_1^1, \theta_3^0, \ldots, \theta_k^0) \\
\theta_3^1 &\sim f(\theta_3 | \theta_1^1, \theta_2^1, \theta_4^0, \ldots, \theta_k^0) \\
&\vdots \\
\theta_k^1 &\sim f(\theta_k | \boldsymbol{\theta}^1_{(-k)})
\end{aligned} \tag{2.7}$$

Note that each draw is conditional on the most recently sampled version of the remaining variables. The above cycle completes a transition from $\boldsymbol{\theta}^0$ to $\boldsymbol{\theta}^1$. Iteration of this algorithm produces the realization $\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, \ldots$ of the Markov chain. Geman and Geman (1984) and Schervish and Carlin (1992) show that under mild regularity conditions (typically aperiodicity and irreducibility) properties (2.6) hold, with the equilibrium distribution $f(\boldsymbol{\theta})$. In a Bayesian context, all full conditional distributions are also conditional on the data $\mathbf{Y}$ so that the equilibrium distribution is the posterior for $\boldsymbol{\theta}$.

Each of the elements of $\boldsymbol{\theta}$ could in fact be a subvector or other group of parameters. Where possible, highly correlated elements are sampled as a group. For example, it can be simple to sample a vector of regression coefficients $\boldsymbol{\beta}$ all at once.

The application of this approach (and MCMC techniques in general) is to run the chain "long enough" so that it has converged. Methods for generating N samples from $f(\boldsymbol{\theta})$ fall between the following two extremes: 1) start N different chains, run each m times, and take the last value from each chain as a sample from $f(\boldsymbol{\theta})$; 2) run one long chain, and after discarding a set of initial values, treat the last N subsequent realizations as the sample. A discussion of the relative merits of these approaches is given in Gelman and Rubin (1992) and Geyer (1992). According to Gelman and Rubin, most practitioners prefer to use several long chains for convergence diagnostics. Additional discussion of convergence diagnostics is given in Sections 3.1 and 5.6.1.

One important aspect of the algorithm, and a large part of its strength is the fact that all sampling is based on full conditional distributions, namely the distribution of each element conditional on all the others. Even when the joint distribution is quite complex, the conditionals may be easy to draw from. Applications where this is true include constrained parameter models, where the range of one parameter may be restricted by the value of another, hierarchical models, in which levels of parameters depend only on those immediately "above" and "below" them, and missing data problems, in which the conditional distributions are simplified when missing data are known.

This last situation (referred to as *data augmentation* - see Tanner and Wong, 1987) is of special interest, since it is used here. Suppose that $\mathbf{Z} = (\mathbf{Y}, \mathbf{Y}')$ denotes the "complete" data, but only $\mathbf{Y}$ is observed. The posterior $f(\boldsymbol{\theta}|\mathbf{Y})$ is typically messy, whereas the posterior $f(\boldsymbol{\theta}|\mathbf{Z})$ based on the (unobserved) complete data are often of a simpler form. The missing data $\mathbf{Y}'$ may be treated as a parameter, and included in a sampling algorithm that samples $\boldsymbol{\theta}$ and $\mathbf{Y}'$, each conditional on the other and the observed data $\mathbf{Y}$. The full conditionals have the form

$$\begin{aligned} f(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Y}') &= f(\boldsymbol{\theta}|\mathbf{Z}) \\ f(\mathbf{Y}'|\boldsymbol{\theta}, \mathbf{Y}) &= f(\mathbf{Y}'|\boldsymbol{\theta}) \end{aligned} \tag{2.8}$$

What is interesting is that the both of these steps are easy. The distribution $f(\boldsymbol{\theta}|\mathbf{Z})$ is of a simpler form than $f(\boldsymbol{\theta}|\mathbf{Y})$, and the $f(\mathbf{Y}'|\boldsymbol{\theta})$ is the sampling distribution of the unobserved data, given $\boldsymbol{\theta}$. Note that generation of $\boldsymbol{\theta}|\mathbf{Z}$ may be accomplished by some iterative scheme, such as subdividing $\boldsymbol{\theta}$ into $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ and replacing

$$f(\boldsymbol{\theta}|\mathbf{Z}) \qquad \text{with} \qquad f(\theta_i|\mathbf{Z}, \boldsymbol{\theta}_{(-i)}), \qquad i = 1, \ldots, k,$$

and using the Gibbs algorithm (2.7).

2.2.2 The Gibbs Sampler for Ordinal Data

The model to be fit is a Bayesian version of the GLM described in Section 2.1, with a probit link, a normal prior for the coefficients $\boldsymbol{\beta}$, and an ordered normal prior for the cutpoints γ. While importance sampling is a possible approach for obtaining the posterior for $(\gamma, \boldsymbol{\beta})$, the Gibbs sampler is easier to extend to a variable selection approach considered later in the thesis. The derivation of conditional distributions for the Gibbs sampler is similar to

Albert and Chib (1993), but with a different parameterization and informative priors. The ability to vary the shape of the prior will make it possible to assess the robustness of the analysis.

The form of the full conditional distributions is simplified by data augmentation, as described in the previous section. Here, the "missing data" are an underlying continuous version of the categorical response. In industrial applications there is quite often such a variable, and an ordered categorical version of it is observed due to cost or time constraints.

It is assumed that for each response Y_i, there are predictors X_i and an unobserved variable Z_i on the continuous scale. The correspondence between Z_i and the J categories of the response is via "cutpoints" $\gamma_0, \gamma_1, ..., \gamma_{J-1}, \gamma_J$, where $-\infty = \gamma_0 < \gamma_1 < ... < \gamma_{J-1} < \gamma_J = \infty$. If

$$\gamma_{j-1} \leq Z_i < \gamma_j,$$

then

$$Y_i = j$$

is observed. This means that Z_i has distribution function

$$\Phi(Z_i - X_i'\boldsymbol{\beta})$$

since from (2.1),

$$\Pr(Z_i < \gamma_j) = \Pr(Y_i \leq j) = \Phi(\gamma_j - X_i'\boldsymbol{\beta}).$$

For a probit link, which uses a standard normal cdf as the link, $Z_i \sim N(X_i'\boldsymbol{\beta}, 1)$.

Full conditional distributions are now derived for arbitrary independent priors on $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, as well as an arbitrary link function. The full conditional distributions are required only up to a proportionality constant, since they are used only for random variable generation. Consequently, the full conditionals are obtainable from the joint distribution of $\mathbf{Y}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{Z}$. This distribution is degenerate, since knowledge of $(\mathbf{Z}, \boldsymbol{\gamma})$ determines $\mathbf{Y}$ exactly. Thus,

$$f(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{Z}, \mathbf{Y}) = f(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{Z})\mathcal{I}(\mathbf{Y}; \mathbf{Z}, \boldsymbol{\gamma}), \tag{2.9}$$

where the indicator function $\mathcal{I}$ is one when $\mathbf{Y}$ agrees with $(\boldsymbol{\gamma}, \mathbf{Z})$, and zero otherwise. Using conditioning, and assuming independence of priors for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ yields:

$$\begin{aligned} f(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{Z}, \mathbf{Y}) &= f(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\gamma}) f(\boldsymbol{\beta}, \boldsymbol{\gamma})\mathcal{I}(\mathbf{Y}; \mathbf{Z}, \boldsymbol{\gamma}) \\ &= f(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\gamma}) f(\boldsymbol{\beta}) f(\boldsymbol{\gamma})\mathcal{I}(\mathbf{Y}; \mathbf{Z}, \boldsymbol{\gamma}). \end{aligned}$$

Note that the distribution of $\mathbf{Z}$ given $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ does not depend on $\boldsymbol{\gamma}$. This follows, since the only dependence of $\mathbf{Z}$ on $\boldsymbol{\gamma}$ is through $\mathbf{Y}$, which is not being conditioned upon in the first term of the expression above. This final simplification gives

$$f(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{Z}, \mathbf{Y}) = f(\mathbf{Z}|\boldsymbol{\beta}) f(\boldsymbol{\beta}) f(\boldsymbol{\gamma})\mathcal{I}(\mathbf{Y}; \mathbf{Z}, \boldsymbol{\gamma}).$$

The full conditional distributions for the Gibbs sampler are immediately available. Their general forms and specific distributions used here are

$$\begin{aligned} f(\boldsymbol{\beta}|\boldsymbol{\gamma},\mathbf{Z},\mathbf{Y}) &\propto f(\mathbf{Z}|\boldsymbol{\beta})f(\boldsymbol{\beta}) \\ &= N((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z},(\mathbf{X}'\mathbf{X})^{-1}) \times N(0,\boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \\ &= N((\boldsymbol{\Sigma}_{\boldsymbol{\beta}}{}^{-1}+\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z},(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}{}^{-1}+\mathbf{X}'\mathbf{X})^{-1}), \end{aligned} \tag{2.10}$$

$$\begin{aligned} f(\boldsymbol{\gamma}|\boldsymbol{\beta},\mathbf{Z},\mathbf{Y}) &\propto f(\boldsymbol{\gamma})\mathcal{I}(\mathbf{Y};\mathbf{Z},\boldsymbol{\gamma}) \\ &= I(\gamma_1 < \ldots < \gamma_{J-1}) \prod_{j=1}^{J-1} N(0,\sigma_{\gamma_j}) \times \mathcal{I}(\mathbf{Y};\mathbf{Z},\boldsymbol{\gamma}) \\ &= \prod_{j=1}^{J-1} N(0,\sigma_{\gamma_j}) \times I(\max\{Z_i|Y_i=j-1\} < \gamma_j < \min\{Z_i|Y_i=j\}), \end{aligned} \tag{2.11}$$

$$\begin{aligned} f(\mathbf{Z}|\boldsymbol{\beta},\boldsymbol{\gamma},\mathbf{Y}) &\propto f(\mathbf{Z}|\boldsymbol{\beta})\mathcal{I}(\mathbf{Y};\mathbf{Z},\boldsymbol{\gamma}) \\ &= N(\mathbf{X}'\boldsymbol{\beta},\mathbf{I})\mathcal{I}(\mathbf{Y};\mathbf{Z},\boldsymbol{\gamma}) \\ &= \prod_{i=1}^{n} N(X_i'\boldsymbol{\beta},1)I(\gamma_{\mathrm{Y_i}-1} < Z_i < \gamma_{\mathrm{Y_i}}). \end{aligned} \tag{2.12}$$

The generic indicator function I is one when its argument is true and zero otherwise. In (2.10) the product of two normals yields a normal distribution. The indicator functions in (2.11) may be combined into a single indicator, which is 1 only when γ_j lies between Z_i elements corresponding to $Y_i = j-1$ and $Y_i = j$ (assuming that there is at least one observation in each outcome group). The truncated distribution for $\mathbf{Z}$ is similar, with each element falling between a pair of cutpoints. Notice that in both (2.11) and (2.12), the elements of $\mathbf{Z}$ and $\boldsymbol{\gamma}$ have independent full conditional distributions, implying that the elements of the vectors may be sampled one at a time.

The practical reason for these choices is that the conditional distributions are all of a form that facilitates sampling. However, the normal priors allow considerable flexibility, and will help illustrate the application of these techniques without diverting too much attention to computational issues. The applications considered here concern orthogonal designs with centered and scaled predictor terms. This implies that the prior mean of $\mathbf{0}$ for $\boldsymbol{\beta}$ is justified, provided there is no prior information regarding the direction in which the effects might act. The experiments considered are primarily screening experiments, so it seems plausible that there would be little prior knowledge about the coefficients, other than the assumption that the effects are not infinite on a transformed scale. As mentioned previously, $\boldsymbol{\beta}$ does not contain an intercept term, meaning that a zero mean for all elements is justifiable. The priors for $\boldsymbol{\gamma}$ are also justifiable, since they allow considerable flexibility, and include two important special cases: A uniform prior on proportions (due to the probit link, and taking $\sigma_{\gamma_j} = 1, j = 1,\ldots J-1$), and a uninformative prior on the transformed scale (taking $\sigma_{\gamma_j} \to \infty, j = 1,\ldots J-1$). By using a parameterization which places all the intercepts in the same vector rather than placing one in $\boldsymbol{\beta}$, separate and independent priors on $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ make more sense.

Thus the Gibbs sampler will sample $\boldsymbol{\beta}$ all at once, and γ and $\mathbf{Z}$ one element at a time. The order of sampling is $\mathbf{Z}$, γ, $\boldsymbol{\beta}$, and all the distributions are easy to sample. If other priors were felt to be appropriate, methods for generating samples from the more complex full conditional distributions are available, and are discussed in Carlin and Gelfand (1991), Gilks and Wild (1992) and Wakefield, Gelfand, and Smith (1991). The algorithm described above is implemented in Ratfor (Kernighan, 1977), a "Rational Fortran" dialect, and for most problems runs quite rapidly. The majority of analysis is done using S-PLUS (Statistical Sciences, 1991).

2.3 Analysis Techniques

The strategy for the Bayesian analysis of industrial data is presented in this section. The content here will either be general or review of existing strategy; the new techniques arising from the application of Bayesian techniques to designed experiments will be introduced via case studies, in Chapter 3. The major task in fitting the models will be variable selection, namely simplification of the model achieved by restricting elements of $\boldsymbol{\beta}$ to be 0. An approach for this is discussed in Section 2.3.1, and model choice criteria used in conjunction with this approach are reviewed in Section 2.3.2. Once the model has been fit, it may be used to optimize the process, as described in Section 2.3.3.

2.3.1 Variable Selection

The basic paradigm for variable selection will be a non-automated stepwise deletion search (but see Chapters 4 and 5 for a more flexible automatic technique). This will typically start with the fitting of the largest model considered, and gradual removal of terms that appear unimportant, with refitting at each step.

It may not be immediately clear why a stepwise search is needed; the design is orthogonal, and in least squares analysis of variance, the estimated effects are uncorrelated and may be examined independently of each other. What is different here is the nonlinear link function, which introduces correlations in the posterior distribution. Hence, deletion (or addition) of a term in a model will change the posterior of the other terms. Each possible model of interest must be compared to the others on some basis, instead of using an ANOVA-like approach of examining the full model only.

The proposed stepwise deletion procedure is carried out by hand due to the lack of stepwise variable selection procedures capable of dealing with infinite estimates. SAS PROC LOGIST has a stepwise variable selection procedure for McCullagh models, but experiences difficulties when the estimates do not converge. While automated versions of this algorithm could be constructed, they are not considered here because the second part of the thesis covers a more flexible automated procedure.

Since the procedure is carried out in a stepwise fashion, a means of identifying "unimportant" terms is required. The approach of Hamada and Wu (1992), who use summaries of the marginal posteriors to suggest which terms are unimportant, will be adopted here.

Some possible summaries of the marginal distributions include the mean, standard deviation, median, other quantiles, and the proportion of the distribution on each side of zero. Roughly speaking, terms with small ratios of posterior mean to standard deviation, medians near zero, quantiles which change sign in the range (say) 5% to 95%, or considerable mass on both sides of zero will be identified as unimportant. The ratio of mean to standard deviation and the proportion of the distribution on one side of zero do not depend on the scaling of the predictors. The fact that the mean, standard deviation, median, and quantiles are not scale invariant is not a concern since the designs are orthogonal, and scaled so that each column in the estimation matrix has variance one.

The iterative procedure begins with fitting the full model. All terms which are clearly unimportant (say with smaller posterior means than standard deviations) are identified, and the model refit without them. Any additional terms that appear unimportant are removed, perhaps using a more strict criterion, and the procedure repeated until all the remaining terms appear important. Early in the procedure, a conservative outlook is taken in dropping terms because the process will be repeated, so unimportant terms which escape removal will be removed at a later iteration. The reason for this repetition is similar to the justification for stepwise model selection. In situations where there are correlations among the estimates, the apparent importance of various terms can change, depending on what other terms are included in the model. By stepping through a sequence of models, the selection procedure is more sensitive to such changes in the importance of the effects, and more likely to arrive at a good model for the data.

Although the marginal summaries are useful for identification of unimportant terms, they are not good overall indicators of the fit of the model to the data. Other non-marginal measures must be used to determine what models appear good, and when to stop the stepwise procedure. Several criteria are considered in the next section, although it will be seen there that the likelihood ratio and penalized versions of it corresponding to approximate Bayesian information criteria are most easily calculated.

2.3.2 Goodness of Fit Criteria

This section reviews some possible criteria for model comparison, and discusses their differences. Suppose models M_0 and M_1 with parameters $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ are to be compared. Model M_0 is a subset of M_1, so $\boldsymbol{\theta}_0$ will be of the same form as $\boldsymbol{\theta}_1$, except for the restrictions imposed by M_0 (typically some elements of $\boldsymbol{\theta}_0$ are 0). Criteria of the form

$$R = \frac{A(\mathbf{Y}, M_0)}{A(\mathbf{Y}, M_1)} \tag{2.13}$$

are considered, where $A(\mathbf{Y}, M)$ is a measure of how well model M describes the data $\mathbf{Y}$. Larger A imply a more appropriate model, so that $R \ll 1$ and $R \gg 1$ suggest support for models M_1 and M_0 respectively. In some cases (the likelihood ratio, for example), $R \gg 1$ is unlikely even when M_0 is preferred. Support for the smaller model M_0 is thus also implied by $R \approx 1$, since there is little evidence to distinguish the models. Since the likelihood defines the structure of the model, it is a component of three criteria considered.

The the likelihood ratio (LR) is obtained by evaluating the likelihood at the MLE $\widehat{\boldsymbol{\theta}}_i$ under model M_i:

$$A(\mathbf{Y}, M_i) = f(\mathbf{Y}|\widehat{\boldsymbol{\theta}}_i, M_i). \tag{2.14}$$

Bayesians average the likelihood over the prior to obtain the Bayes factor (BF) from (2.13), with

$$A(\mathbf{Y}, M_i) = \int f(\mathbf{Y}|\boldsymbol{\theta}_i, M_i) f(\boldsymbol{\theta}_i|M_i) d\boldsymbol{\theta}_i = f(\mathbf{Y}|M_i). \tag{2.15}$$

A third criterion, the posterior Bayes factor (PBF), suggested by Aitkin (1990), compares likelihoods averaged over the posterior:

$$A(\mathbf{Y}, M_i) = \int f(\mathbf{Y}|\boldsymbol{\theta}_i, M_i) f(\boldsymbol{\theta}_i|M_i, \mathbf{Y}) d\boldsymbol{\theta}_i. \tag{2.16}$$

Each of these methods has some justification. The LR compares models at their best, in the sense that the MLE's provide the best fit to the data within a given model. The BF may be interpreted as an odds ratio of prior and posterior probabilities, since

$$BF = \frac{f(\mathbf{Y}|M_0)}{f(\mathbf{Y}|M_1)} = \frac{f(M_0|\mathbf{Y})f(\mathbf{Y})}{f(M_0)} \Big/ \frac{f(M_1|\mathbf{Y})f(\mathbf{Y})}{f(M_1)} = \frac{f(M_0|\mathbf{Y})}{f(M_1|\mathbf{Y})} \times \frac{f(M_1)}{f(M_0)}. \tag{2.17}$$

That is, the Bayes factor compares the posterior probabilities of M_0 and M_1, weighted by their prior probabilities. If the models are a priori equiprobable, then the BF picks the most probable one given the data. The PBF may be interpreted as a comparison of the ability of the models to predict the observed data as if they were future data. We see this since the predictive distribution of some future observation $\mathbf{Y}_F$ given the current data and model M_i is

$$f(\mathbf{Y}_F|M_i, \mathbf{Y}) = \int f(\mathbf{Y}_F|\boldsymbol{\theta}_i, M_i, \mathbf{Y}) f(\boldsymbol{\theta}_i|M_i, \mathbf{Y}) d\boldsymbol{\theta} = \int f(\mathbf{Y}_F|\boldsymbol{\theta}_i, M_i) f(\boldsymbol{\theta}_i|M_i, \mathbf{Y}) d\boldsymbol{\theta}. \tag{2.18}$$

Note that in obtaining the last expression in (2.18), the assumption is made that given a model M_0, parameters $\boldsymbol{\theta}_0$, and the data $\mathbf{Y}$, the future observation does not depend on the data $\mathbf{Y}$. Substitution of $\mathbf{Y}_F = \mathbf{Y}$ in (2.18) yields an assessment of how well the model predicts data with the same sufficient statistics as the observed data.

Asymptotic theory has been applied to all three of these criteria, to obtain test sizes for the LR, and to surmount difficulties in calculating the BF and PBF. In the case of LR, asymptotic theory states that under certain regularity conditions (see, for example Wilks (1962)),

$$-2\log(LR) = -2(\log f(\mathbf{Y}|\widehat{\boldsymbol{\theta}}_0, M_0) - \log f(\mathbf{Y}|\widehat{\boldsymbol{\theta}}_1, M_1)) \xrightarrow{d} \chi^2_\nu \tag{2.19}$$

with

$$\nu = \mathrm{df}(M_1) - \mathrm{df}(M_0).$$

Formal significance tests may be based on this distribution.

Table 2.1 Calibration of LR, BF, PBF

LR Critical Values						Evidence against M_0	BF	PBF
α	$\nu = 1$	$\nu = 2$	$\nu = 5$	$\nu = 10$	$\nu = 20$	moderate	3.0e-1	–
0.05	1.4e-1	5.0e-2	3.9e-3	1.1e-4	1.5e-7	strong	1.0e-1	5.0e-2
0.01	3.6e-2	1.0e-2	5.3e-4	9.1e-6	7.0e-9	very strong	1.0e-2	1.0e-2
0.001	4.4e-3	1.0e-3	3.5e-5	3.8e-7	1.4e-10	decisive	<1.0e-2	1.0e-3

In the case of the BF and PBF, evaluation of the necessary integrals is often so difficult and time consuming (even with the aid of numeric techniques) that large sample approximations have been derived to avoid calculation. For the exponential family, and priors that remain fixed as sample size increases, Schwarz (1978) illustrates that the BF can be approximated as

$$-2\log(BF) = -2\log(LR) - \nu \log n. \tag{2.20}$$

Aitkin (1990) points out that (2.20) should apply in general, provided that the likelihood is roughly normal (that is, the sample size is large relative to the number of parameters) and the priors vary gradually relative to the likelihood. Under similar conditions, he gives the approximation

$$-2\log(PBF) = -\log(LR) - \nu \log 2. \tag{2.21}$$

Another approximation for the BF is obtained under the assumption that the priors vary with sample size. Smith and Spiegelhalter (1980) assume that the prior for $\boldsymbol{\theta}_1$ is in some "local neighborhood" of the $\boldsymbol{\theta}_0$ prior, and that as the sample size increases, the neighborhood shrinks toward the $\boldsymbol{\theta}_0$ prior. Under these conditions (specifically for normal linear models, but the comments of Aitken mentioned above apply), a form similar to Akaike's AIC is obtained:

$$-2\log(BF) = -2\log(LR) - \nu\tfrac{3}{2}. \tag{2.22}$$

Note that for the AIC, 3/2 is replaced by 2.

While these approximations may not be good for the small samples encountered in many industrial problems, they do indicate the overall behavior of the three criteria. As Spiegelhalter and Smith (1982) point out, "for large experiments evidence at a very high significance level is required for the Bayes factor to favor strongly the more complex hypothesis". The LR behaves differently, favoring more complex models as the sample size increases. The PBF is a compromise between the two, since it has a penalty that does not depend on sample size n. That the PBF is a compromise between the LR and the BF is also evident from its construction. While it still averages the likelihood like a BF, the averaging places the most mass near the MLE's (for priors that do not strongly contradict the data).

The calibration of these statistics is obtained by different approaches. For the likelihood ratio LR, formal testing is based upon the chi-square approximation, which indicates what an unusual ratio is. For the BF, Jeffreys (1961, Appendix B) proposes "order of magnitude" interpretations. Aitken (1990) proposes similar interpretation of values for the PBF. These values are compared on the original ratio scale in Table 2.1. For small ν the calibrations of

LR, BF, and PBF are directly comparable. For larger ν there are substantial differences. This is because the LR makes adjustments for ν in the critical values, while adjustments for ν (and n in the case of the BF) are made implicitly in the calculation of BF and PBF. Although Aitkin's values for the PBF are slightly different from Jefferys' suggestions for the BF, Aitkin indicates that the PBF should be interpreted in a fashion similar to the BF.

Some features of the BF are worth noting. It will be defined only for proper (i.e. informative) priors on $\boldsymbol{\theta}$, since otherwise $f(\boldsymbol{\theta}|M_i)$ in (2.15) involves an arbitrary constant, meaning that the ratio (2.13) will also be arbitrary. The fact that the BF uses likelihoods averaged over priors means that it can be quite sensitive to priors. The PBF, on the other hand is much less sensitive, and the LR does not utilize priors.

For the ordinal regression problem with data $\mathbf{Y} = Y_1, \ldots, Y_n$ and predictor matrix $\mathbf{X}$, the likelihood is based on the model (2.1), and has the form

$$\prod_{i=1}^{n} \Pr(Y_i|X_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} [\Phi(\gamma_{Y_i} - X_i'\boldsymbol{\beta}) - \Phi(\gamma_{Y_i} - X_i'\boldsymbol{\beta})]. \tag{2.23}$$

The computational complexity in evaluating the BF and PBF will depend on both the sample size and the dimension of the parameter space $(\boldsymbol{\gamma}, \boldsymbol{\beta})$. Monte Carlo integration techniques (see Section 2.2.1) were considered for calculation of the BF and PBF. A version without importance sampling averages the likelihood $f(\mathbf{Y}|\boldsymbol{\gamma}, \boldsymbol{\beta})$ over draws $\{\boldsymbol{\gamma}^i, \boldsymbol{\beta}^i\}_{i=1}^{N}$ from the prior $f(\boldsymbol{\gamma}, \boldsymbol{\theta})$. This is computationally inefficient, since the prior and posterior differ considerably. An importance sampling approach was attempted, using draws (for $\boldsymbol{\beta}$) from a $N(\overline{\boldsymbol{\beta}}, \overline{\Sigma})$ distribution, where $\overline{\boldsymbol{\beta}}$ is the posterior mean of $\boldsymbol{\beta}$ and $\overline{\Sigma}$ is the posterior covariance, both based on a Gibbs sample. For $\boldsymbol{\gamma}$, independent truncated normals were used with appropriate posterior means and variances. This approximation is used because it is necessary to evaluate the density from which the samples are drawn to obtain the importance weights. The true posterior cannot be evaluated because the normalizing constant is not known. The importance sampling approach was also intractable, because it required enormous samples. For example in the handle data (see Section 3.2) importance sampling was used to generate 500,000 Monte Carlo samples for a model with 11 predictors and 5 cutpoints. Two different runs produced results that were orders of magnitude different.

The unavailability of exact numeric results means that other information criteria or approximations must be used. The second half of the thesis deals with an algorithm for calculating posterior probabilities of models, so here the strategy will be to use simple approximations to measure the fit of a model. These measures will consist of the likelihood ratio and penalized versions of it (BIC, AIC, approximate PBF) that approximate the BF or PBF. While the asymptotic results are unlikely to hold, these criteria will still be useful in identifying models that fit the data well.

Evaluation of these criteria simply requires maximization of the likelihood over the parameter space. When the MLE's exist, this evaluation is straightforward, and commonly available software, such as SAS PROC LOGISTIC (SAS, 1991) or McCullagh's (1988) PLUM package may be used. When the MLE's diverge, the maximum of the likelihood is still finite (as mentioned in Section 2.1), and can be used to asses the fit of the model.

Unfortunately, both of these packages halt their iterative procedure before the likelihood has converged. Since neither package is easily modified to continue iteration, alternative means of maximizing the likelihood must be used.

While general maximization packages could be used, the available Gibbs sampling software can also maximize the likelihood. To obtain rough values of the maximum likelihood, a diffuse prior is used, and the likelihood (2.23) evaluated for each of the posterior samples. The samples are from the likelihood, so the largest value of the likelihood evaluated over the samples will be a rough estimate of the true maximum. The divergent estimates mean that the likelihood is very flat, with asymptotes at $\pm\infty$ in the direction of non-existent estimates. The flatness of the likelihood means that many samples from it will give a value of the likelihood near its maximum.

2.3.3 Process optimization

Once a model has been identified, it may be used to optimize the process. If possible, additional data should be gathered to ensure more realistic estimation of the model identified by the variable selection step. Otherwise, this model will tend to fit the original data set too well, reducing the quality of the prediction. In practice, additional data are often not available, and the same model must be used for optimization. This will be the case for both of the examples given in the next section. Fortunately the optimization procedure is the same whether additional data are gathered or not.

The goal of optimization is to find settings of the predictors which yield high probability content in desired outcome groups. Typically there will be a single category that is the "target", whether it is at one extreme of the categories (as in Section 3.1), or a category somewhere in the middle (as in Section 3.2). For any given setting of predictors X and parameters $(\boldsymbol{\gamma}, \boldsymbol{\beta})$, the model

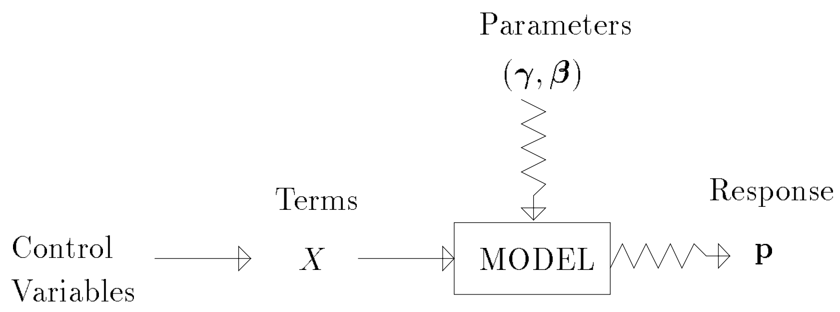
$$\Pr(Y_i \leq j) = \Phi(\gamma_j - X_i'\boldsymbol{\beta}) \qquad \text{for} \qquad j = 1, ..., J-1 \tag{2.1}$$

may be used to obtain a vector $\mathbf{p}$ of (non-cumulative) probabilities for the outcome groups of response Y, with elements

$$p_j = \Pr(Y = j) = \Phi(\gamma_j - X'\boldsymbol{\beta}) - \Phi(\gamma_{j-1} - X'\boldsymbol{\beta}). \tag{2.24}$$

This relationship between $X, \boldsymbol{\gamma}, \boldsymbol{\beta}$ and $\mathbf{p}$ is illustrated in Figure 2.1. In this figure, straight lines indicate deterministic relations, while distributions are involved for jagged lines. The "control variables" are variables actually set during the experiment, while the vector X of terms may also include other constructed variables such as interactions. The straight lines for the control variables and X indicate that once the control variables are set, no variation is transmitted to the model. The uncertainty about parameters $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ (i.e. the variation in their posteriors) is represented by the jagged line leading to the model. This uncertainty is transferred to $\mathbf{p}$ via the model (2.1), and so $\mathbf{p}$ has a posterior distribution. To avoid possible confusion due to a probability having a posterior, $\mathbf{p}$ will be referred to as vector of

Figure 2.1 Optimization without noise variables



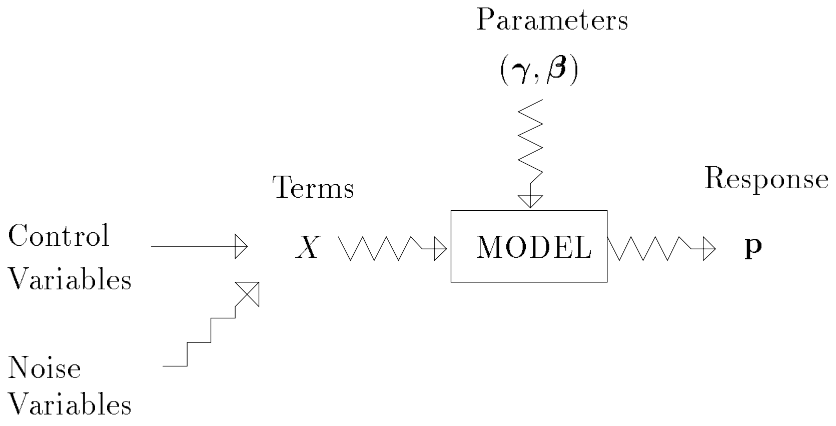
proportions with a posterior. This posterior may be "adjusted" by choosing different levels of the control variables. The desired posterior of $\mathbf{p}$ will be a distribution of proportions near 1 (for the desired category) and 0 (for all other categories) with as little variation as possible. This goal corresponds to Taguchi's goal of an on-target response with minimum variance.

A simple example serves to clarify this. Suppose A and B are control variables, and the model contains terms A, B, and an AB interaction. Assume that A and B are fixed. If a sample $\{\boldsymbol{\gamma}_i, \boldsymbol{\beta}_i\}_{i=1}^N$ from the posterior $f(\boldsymbol{\gamma}, \boldsymbol{\beta}|Y)$ is available, then a sample from the posterior $f(\mathbf{p}|Y)$ is easily obtained, by calculation of $\mathbf{p}_i$ from (2.24) for each of $(\boldsymbol{\gamma}_i, \boldsymbol{\beta}_i)$. Optimization will consist of repetition of this calculation for different fixed levels of A and B.

In robust design experiments, there is an additional challenge to the optimization process. The predictors are divided into two groups: those which are controllable under regular production (control variables) and those which are expensive or impossible to control under production (noise variables). A special off-line experiment is carried out in which control and noise variables are set at fixed levels. The response is modeled as a function of both the control and noise variables (here with model (2.1)), and the model used to choose control settings that desensitize the response to the variation in the noise variables. Modeling the response directly was proposed by Welch, Yu, Kang and Sacks (1990) and named the response model approach by Shoemaker, Tsui and Wu (1991). Once a model has been selected, the noise variables are assumed to follow a specific distribution. Figure 2.2 illustrates the relationship between the control variables, noise variables, $\boldsymbol{\gamma}, \boldsymbol{\beta}$, and response $\mathbf{p}$. The distribution of $\mathbf{p}$ now has two different sources of variability - the posterior for $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and the distribution of the noise variables. There may be considerable opportunity to reduce the variability in $\mathbf{p}$ due to the noise variables. This depends on the presence of large interactions between control and noise variables. By choosing different levels of the noise variables, the effect of the uncontrollable noise variables on $\mathbf{p}$ can be reduced.

To continue the simple example, suppose that an additional noise variable W is studied,

Figure 2.2 Optimization with noise variables



and the model consists of terms A, B, W, AB, AW, and BW. In the off-line experiment, levels of A, B, and W are set at fixed levels. The model (2.1) is fit to the data and posteriors obtained. Since the goal of the optimization is to optimize the process under a production environment, the noise variable W is then assumed to follow some distribution (say $N(0,1)$). The variables A and B are still set at some fixed levels. The additional variation in W is transferred to $\mathbf{p}$, via the model

$$\begin{aligned}\Pr(Y \leq j) &= \Phi(\gamma_j + [A\beta_A + B\beta_B + W\beta_W + AB\beta_{AB} + AW\beta_{AW} + BW\beta_{BW}]) \\ &= \Phi(\gamma_j + [A\beta_A + B\beta_B + AB\beta_{AB} + W(\beta_W + A\beta_{AW} + B\beta_{BW})]).\end{aligned}$$

By choosing the levels of A and B, it may be possible to make the term $\beta_W + A\beta_{AW} + B\beta_{BW}$ small, reducing the overall variation in $\mathbf{p}$.

Noise variables are easily incorporated into the calculation of the distribution of proportions. For each $(\boldsymbol{\beta}_i, \gamma_i)$ sampled from the posterior, a realization of the noise variables is sampled from their distribution (here $W_i \sim N(0,1)$), and (2.1) used to calculate $\mathbf{p}_i$. Repetition of this for the entire sample yields a sample from the distribution of proportions for a given setting of the control variables.

It should be emphasized that the variation due to noise variables is treated in the same fashion as parameter uncertainty, since both are represented by distributions. This flexibility to accommodate both model uncertainty and external variation is another advantage the Bayesian approach offers in analyzing such data.

Chapter 3

Case Studies

In this chapter, two industrial data sets are used to illustrate the analysis of data via the posterior samples. Since exposition rather than exhaustive analysis is the objective, different approaches and problems will be highlighted in each case. Two basic tasks will be common to the industrial situation: model fitting and optimization. The model fitting will include empirical methods for assessing convergence of the Gibbs algorithm, variable selection techniques, and plots for examining the posteriors and their robustness to different priors. Optimization topics will include Taguchi's robust design, on-target and larger the better problems, interpretation using posterior means and full distributions, and projections as interpretive tools.

3.1 A Foam Molding Experiment

The data, which was originally analyzed by Jinks (1987), arise from an experiment designed to reduce voids in an urethane-foam product. The response consists of three levels (very good, acceptable, needs repair), while all the design variables are at two levels, (-1,1). The design is a fractionated eight run control array crossed with a four run noise array. The factors are shown in Table 3.1. At each level of the control and noise factors, ten parts are classified into one of the three categories, yielding a total of $8 \times 4 \times 10 = 320$ observations. These observations and the design are shown in Table 3.2

The full model consists of main effects for factors A-I, an HI interaction, and control by noise interactions between factors A-G and H, I. In total, there are 24 terms plus two cutpoints which define the three ordered categories. A model with diffuse priors for $\boldsymbol{\gamma}$, and a $N(\mathbf{0}, \mathbf{I})$ prior for $\boldsymbol{\beta}$ is used. Other variances could be used, and robustness to this choice is desirable - the next example considers this issue in more detail.

Before the posteriors are analyzed, an empirical method for assessing convergence is presented. Several long chains using different starting values $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)$ are run and observed to see how long they take to reach the same distribution. In Figure 3.1, the behavior of four chains with starting values $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) = \{(\mathbf{0}, (-5, -5)), (\mathbf{0}, (-5, 5)), (\mathbf{0}, (5, 5)), (\mathbf{0}, (0.1, 0.1))\}$

Table 3.1 Factors in the foam molding experiment

	Control Factors		Noise Factors
A	Shot Weight (185 / 250)	H	Shift (second / third)
B	Mold Temperature (70° F / 120° F)	I	Shell quality (good / bad)
C	Foam Block (use / do not use)		
D	RTV Insert (use / do not use)		
E	Vent Shell (vented / unvented)		
F	Spray Wax Viscosity (2:1 / 4:1)		
G	Tool Elevation (level / elevated)		

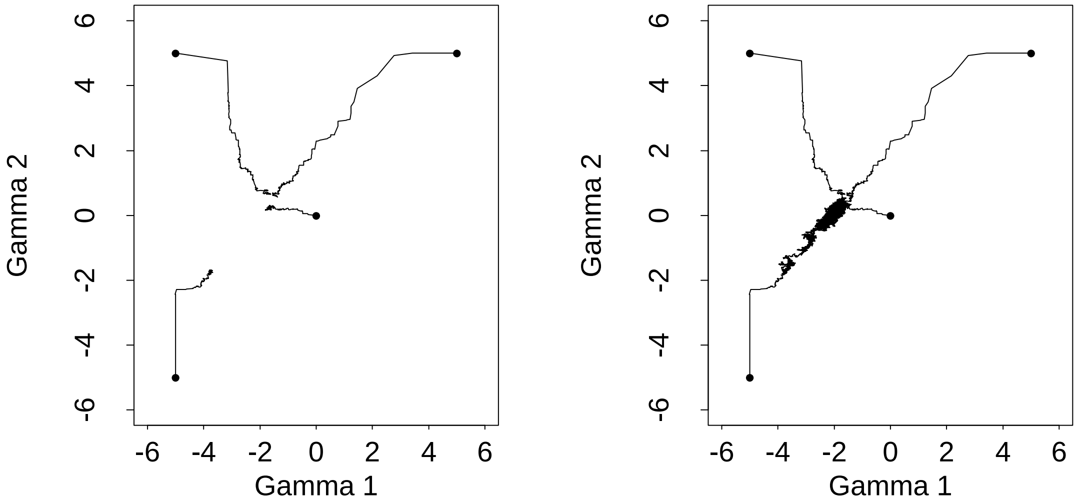
Table 3.2 Foam molding data, with frequencies for good (I), ok (II) and poor (III)

						H		-1			1			-1			1	
						I		-1			-1			1			1	
A	B	C	D	E	F	G	I	II	III	I	II	III	I	II	III	I	II	III
-1	-1	-1	-1	-1	-1	-1	3	6	1	6	4	0	1	4	5	0	10	0
-1	-1	-1	1	1	1	1	0	3	7	3	4	3	0	6	4	0	7	3
-1	1	1	-1	-1	1	1	0	0	10	0	1	9	0	0	10	0	0	10
-1	1	1	1	1	-1	-1	0	0	10	0	10	0	0	3	7	0	9	1
1	-1	1	-1	1	-1	1	3	5	2	3	7	0	3	5	2	1	6	3
1	-1	1	1	-1	1	-1	2	8	0	4	5	1	0	5	5	1	5	4
1	1	-1	-1	1	1	-1	2	7	1	2	5	3	2	7	1	1	6	3
1	1	-1	1	-1	-1	1	0	4	6	1	7	2	0	4	6	0	3	7

is illustrated. The first plot illustrates that after 250 iterations, the sequences have not converged. The second plot shows the same chains after convergence (which occurs after about 500 observations). Note that $\boldsymbol{\beta}_0 = \mathbf{0}$ is used in all four cases. Study of the behavior of the convergence of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ indicated that $\boldsymbol{\gamma}$ tends to move about the parameter space quite slowly, and hence the starting values of $\boldsymbol{\beta}$ are of less concern. The form of the full conditional distribution (2.11) for $\boldsymbol{\gamma}$ and the large sample size cause this slow convergence, since γ_1 and γ_2 are restricted to lie between Z_i's corresponding to Y_i's in categories I, II, and III. There are 38, 156, and 126 responses in the three categories, respectively. Random variation in the Z_i's will leave very narrow regions for sampling $\boldsymbol{\gamma}$ at each cycle, and thus convergence will be slow. This suggests that a single long run of the Gibbs sampler will be most efficient, and the informal diagnostics used above suggest the following strategy: The last 10,000 samples will be used from a single long chain of 11,000 cycles with starting values $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) = (\mathbf{0}, (0.1, 0.1))$.

The first analysis task is to identify the factors with the largest effects, via the stepwise deletion approach outlined in Section 2.3.1. The posteriors obtained via the Gibbs sampling algorithm are quite close to normal; hence a numeric summary of the marginal posteriors

Figure 3.1 Convergence of γ, 250 and 5000 iterations



is given (in Table 3.3) rather than plots. These summaries of the posteriors are used to identify terms which could be dropped from the model. In this case, the variables DH, EH, FH, GH, AI, BI, CI, FI, GI are first dropped, and the model refit. Several of the columns of Table 3.3 support this decision. The means are smaller than corresponding standard deviations, and the quantiles and zero splits indicate that the centers of the posteriors lie near zero. The next iteration leads to the removal of the terms HI, BH, CH, DI, and D, leaving a model with the terms A, B, C, E, F, G, H, I, AH, and EI. The posteriors are qualitatively very similar to those given in Table 3.3, and are not summarized here.

Various information criteria (discussed in Section 2.3.2) which compare the two submodels to the full model are given in Table 3.4. These criteria are all based on the log likelihood ratio (LR), with various penalties. The values given are all of the form -2log(LR), relative to a baseline model containing all 31 predictors and two intercepts. The two models considered in the backwards deletion algorithm have 12 and 17 parameters and are marked with an asterisk. Both the LR and the BIC suggest that the 12 parameter model is a reasonable simplification of the full 26 parameter model. The likelihood ratio test has a p-value of 0.088 for making the reduction to 12 parameters, and the BIC attains its lowest value for this model. The lighter penalties used for the AIC and PBF mean that they prefer models with more terms. Since a parsimonious model is desired, the model with terms A, B, C, E, F, G, H, I, AH, and EI will be used. Note that even a model with 31 predictors and two cutpoints cannot explain all the variation. This due to the variation in the response under replication.

Other models are included in Table 3.4, because there is enough data to guarantee the existence of the MLE for this problem. Consequently, the SAS variable selection routine

Table 3.3 Marginal posterior summaries and p-values for scored ANOVA, foam experiment

			Quantiles							**Zero Split**		**Anova**
	mean	s.d.	0.01	0.05	0.25	0.5	0.75	0.95	0.99	%<0	%>0	p-val
A	-0.75	0.19	-1.15	-1.04	-0.88	-0.77	-0.62	-0.43	-0.34	1.000	0.000	0.000
B	0.91	0.19	0.49	0.58	0.77	0.92	1.04	1.20	1.31	0.000	1.000	0.000
C	0.62	0.19	0.22	0.30	0.48	0.64	0.75	0.92	1.01	0.000	1.000	0.006
D	-0.24	0.19	-0.65	-0.54	-0.38	-0.26	-0.10	0.08	0.16	0.866	0.134	0.019
E	-0.58	0.19	-1.00	-0.88	-0.72	-0.60	-0.44	-0.27	-0.18	1.000	0.000	0.003
F	0.57	0.19	0.16	0.25	0.43	0.58	0.70	0.87	0.98	0.000	1.000	0.006
G	0.75	0.19	0.35	0.43	0.61	0.76	0.89	1.05	1.17	0.000	1.000	0.000
H	-0.45	0.17	-0.91	-0.76	-0.56	-0.43	-0.32	-0.20	-0.13	1.000	0.000	0.000
I	0.37	0.16	0.08	0.15	0.25	0.35	0.47	0.68	0.80	0.000	1.000	0.000
HI	0.19	0.07	0.01	0.06	0.14	0.19	0.24	0.31	0.36	0.006	0.994	0.019
AH	0.43	0.17	0.11	0.18	0.30	0.41	0.54	0.74	0.90	0.000	1.000	0.001
BH	-0.20	0.17	-0.67	-0.51	-0.31	-0.18	-0.07	0.05	0.12	0.896	0.104	0.830
CH	-0.24	0.17	-0.70	-0.55	-0.35	-0.22	-0.11	0.00	0.08	0.945	0.055	0.668
DH	-0.03	0.17	-0.35	-0.28	-0.16	-0.04	0.09	0.29	0.43	0.590	0.410	0.003
EH	0.14	0.17	-0.18	-0.10	0.02	0.13	0.26	0.46	0.60	0.218	0.782	0.830
FH	-0.03	0.17	-0.48	-0.34	-0.14	-0.01	0.10	0.23	0.30	0.520	0.480	0.033
GH	-0.10	0.17	-0.56	-0.41	-0.21	-0.08	0.03	0.15	0.22	0.682	0.318	0.284
AI	-0.03	0.16	-0.46	-0.34	-0.13	-0.01	0.08	0.19	0.26	0.528	0.472	0.134
BI	-0.03	0.16	-0.32	-0.25	-0.15	-0.05	0.07	0.28	0.40	0.620	0.380	0.011
CI	0.11	0.16	-0.18	-0.11	-0.01	0.09	0.22	0.42	0.54	0.270	0.730	0.830
DI	-0.18	0.16	-0.60	-0.49	-0.28	-0.16	-0.06	0.04	0.11	0.892	0.108	0.830
EI	-0.36	0.16	-0.80	-0.67	-0.46	-0.33	-0.24	-0.14	-0.07	0.999	0.001	0.003
FI	0.12	0.16	-0.17	-0.10	0.01	0.10	0.22	0.43	0.54	0.230	0.770	0.521
GI	0.09	0.16	-0.20	-0.13	-0.03	0.06	0.19	0.40	0.52	0.325	0.675	0.199
γ_1	-2.08	0.21	-2.53	-2.42	-2.22	-2.10	-1.96	-1.70	-1.54	1.000	0.000	
γ_2	-0.03	0.21	-0.41	-0.32	-0.18	-0.06	0.11	0.35	0.42	0.632	0.368	

Table 3.4 Model choice criteria for various models, foam data

Model[a]	df	log LR	AIC	BIC	log PBF
γ	2	184.66	136.66	46.22	75.69
γ,B	3	150.49	104.49	17.82	59.30
γ,B,G	4	126.42	82.42	-0.49	47.96
γ,A,B,G,	5	107.04	65.04	-14.10	38.96
γ,A,B,G,H	6	95.18	55.18	-20.18	33.73
γ,A,B,G,H,AH	7	84.33	46.33	-25.27	29.00
γ,A,B,G,H,I,AH	8	73.41	37.41	-30.42	24.23
γ,A,B,E,G,H,I,AH	9	62.53	28.53	-35.53	19.48
γ,A,B,C,E,G,H,I,AH	10	50.02	18.02	-42.27	13.92
γ,A,B,C,E,F,G,H,I,AH	11	34.00	4.00	-52.53	6.60
γ,A,B,C,E,F,G,H,I,AH,EI*	12	21.59	-6.41	-59.16	1.09
γ,A,B,C,E,F,G,H,I,AH,EI,HI	13	15.98	-10.02	-59.01	-1.02
γ,A,B,C,E,F,G,H,I,AH,DH,EI,HI	14	11.73	-12.27	-57.49	-2.45
γ,A,B,C,D,E,F,G,H,I,AH,BH,CH,DI,EI,HI*	17	6.84	-11.16	-45.08	-2.82

[a] Models fit by the non-automatic procedure are marked with an asterisk.

in PROC LOGISTIC can be run on the data. A stepwise approach starting from the null model was used, with significance levels of 0.05 for entering and leaving the model. All models except the one in the last row of Table 3.4 were considered. The stepwise procedure decided to select the model with 14 terms, a slightly more complex model than what will be used here for optimization.

Table 3.3 also gives p-values for a scored ANOVA, with the scores (1,2,3). The ANOVA approach also identifies D, DH, FH, BI as significant, even though none of the likelihood-based criteria or marginal posterior summaries suggest that these terms are important. The agreement between other methods suggests that the problem may be in the scored ANOVA approach. One possible explanation is that scored ANOVA tends to under-estimate the residual variance because of the non-continuous nature of the response. This under-estimation gives higher significance levels than are realistic.

Other methods for exploring the posteriors are recommended; for example, plots of joint distributions and measures of correlation are of interest. In this case, posterior correlations are quite large for some coefficients in the full model, but when the reduced model is fit, the correlations are much smaller.

Consider now the optimization of the process using the fitted model. Here, the goal is to find settings which yield high probability content in certain outcome groups. More accurately (see Section 2.3.3), the goal is to identify settings of the control variables A, B, C, E, F, G which yield a favorable posterior for the vector of outcome proportions $\mathbf{p}$. Following the advice in Section 2.3.3, a "favorable" posterior will have proportions near 1 in the good category, and near 0 in the remaining two. For illustrative purposes, assume that the noise variables $(H, I) \sim N(\mathbf{0}, \mathbf{I})$ under production, although in practice this assumption would be based on expert knowledge. The optimization will be relatively simple - only ± 1 levels of the six control factors are considered.

To identify promising settings, consider $E(\mathbf{p}|A, B, C, E, F, G)$, with the expectation taken over the posterior of $\mathbf{p}$ and noise variables H and I. The expected proportions for the $2^6 = 64$ different control settings are plotted in Figure 3.2. Only two components of $\mathbf{p}$ are plotted, since the three proportions must sum to one. Optimal settings will have the most mass on Pr(Good), and most of the remaining mass on Pr(OK) - namely those near the lower right of the plot and near the diagonal line $\Pr(\text{Good}) + \Pr(\text{OK}) = 1$. In this case, there appears to be a clear winner - the setting labeled 10, with levels $(A, B, C, E, F, G) = (+, -, -, +, -, -)$ and expected proportions of 80% good parts and 20% ok parts. Examination of the entire distribution of $\mathbf{p}$ for setting 10 reveals that most mass lies on the diagonal, so $\sqrt{\text{Var}(\Pr(\text{Good}))} = 0.0873$ provides a good summary of variation. Because one setting is the clear winner, discussion of other interpretive tools for optimization will be postponed until the next example.

3.2 An Injection Molding Experiment

This experiment, which is analyzed by Steinberg and Bursztyn (1993, 1994), was conducted to improve the quality of injection molded plastic handles. The goal of the experiment was to produce "on target" parts, in contrast to the "lower the better" experiment in Section 3.1. The response variable is the "amount of material", ranging from one (too little) to seven (too much), with a target value of four. No values of seven were actually recorded from the data, so the analysis is for the six observed response categories. The original data set consisted of 17 control factors and two noise factors. For illustrative purposes, a quarter fraction of the experiment is analyzed, with no noise effects (the noise factors were set as follows: mold = 2, water temperature = 1). The quarter fraction is a $2^{(17-12)}$ fractional factorial design, with 32 observations. The factors are labeled A-Q, and are grouped as follows: A-D are process temperatures, E-G stroke counts, H-L pressures, and N-Q are process times. All factors are at two levels $(-1, 1)$ in an orthogonal design. The data to be analyzed are shown in Table 3.5.

The Gibbs algorithm converged much more rapidly for this problem, due to the smaller number of observations, which implies more freedom in the full conditional distributions for γ (see the foam experiment for discussion). A single chain of 10250 cycles was generated, and the last 10000 values used.

The model fitting strategy is the same as in the previous example and only a brief summary is given. A saturated model with 31 effects was first considered, and all unimportant terms dropped in groups. The final model chosen is the one with main effects A, C, E, F, I, O, and an AO interaction; the magnitudes of the A and AO terms were not large, but were included in the model. Information measures based on the likelihood were not

Figure 3.2 Expected proportions of good and ok parts at different settings of control factors A, B, C, E, F, G for the foam experiment.

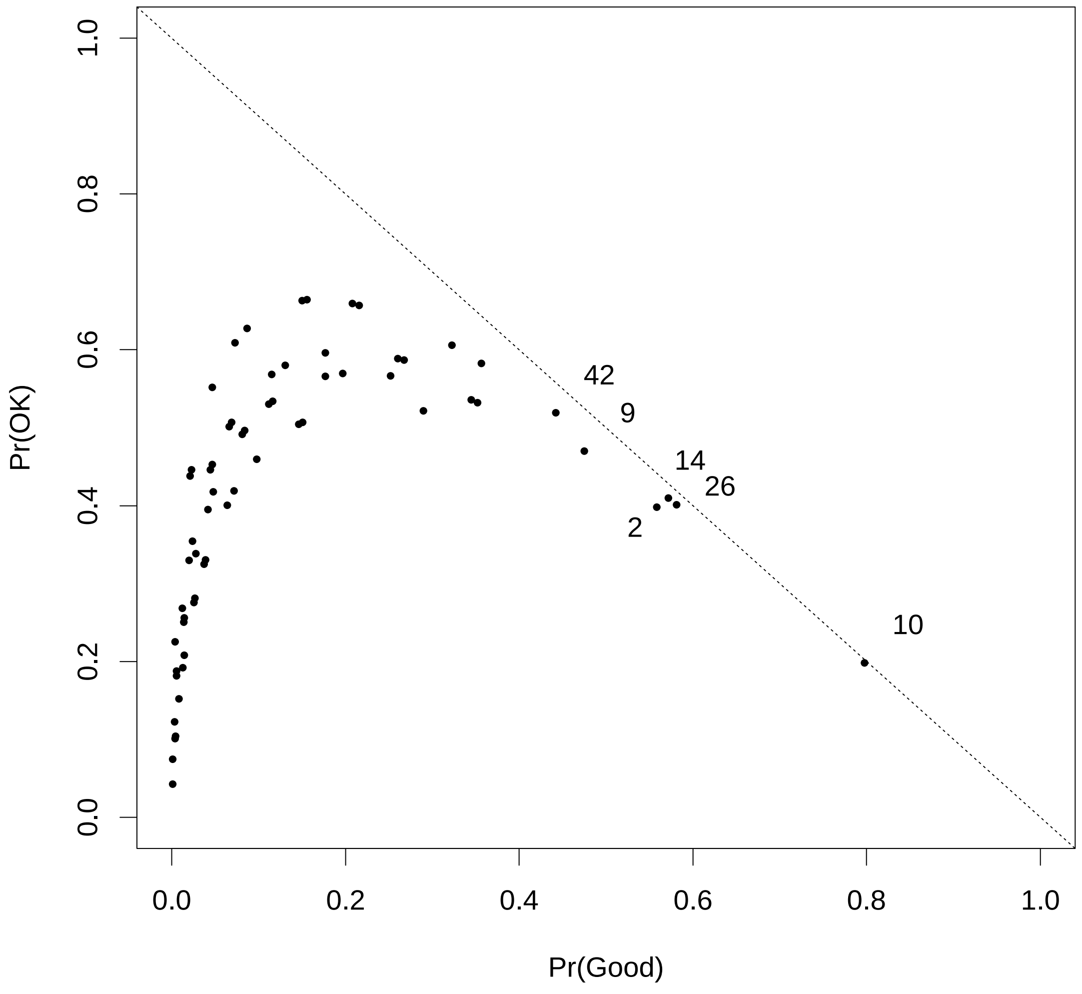


Table 3.5 Handle data

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	Y
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1
-1	-1	-1	-1	-1	-1	-1	-1	1	-1	1	1	1	1	1	1	1	1
-1	-1	-1	1	1	1	1	1	-1	1	-1	-1	-1	-1	1	-1	1	4
-1	-1	-1	1	1	1	1	1	1	1	1	1	1	1	-1	1	-1	5
-1	-1	1	-1	1	-1	-1	1	-1	1	-1	-1	1	1	-1	1	-1	2
-1	-1	1	-1	1	-1	-1	1	1	1	1	1	-1	-1	1	-1	1	2
-1	-1	1	1	-1	1	1	-1	-1	-1	-1	-1	1	1	1	1	1	4
-1	-1	1	1	-1	1	1	-1	1	-1	1	1	-1	-1	-1	-1	-1	5
-1	1	-1	-1	1	1	1	-1	1	1	-1	1	1	-1	-1	-1	1	5
-1	1	-1	-1	1	1	1	-1	-1	1	1	-1	-1	1	1	1	-1	4
-1	1	-1	1	-1	-1	-1	1	1	-1	-1	1	1	-1	1	-1	-1	1
-1	1	-1	1	-1	-1	-1	1	-1	-1	1	-1	-1	1	-1	1	1	1
-1	1	1	-1	-1	1	1	1	1	-1	-1	1	-1	1	-1	1	1	5
-1	1	1	-1	-1	1	1	1	-1	-1	1	-1	1	-1	1	-1	-1	4
-1	1	1	1	1	-1	-1	-1	1	1	-1	1	-1	1	1	1	-1	4
-1	1	1	1	1	-1	-1	-1	-1	1	1	-1	1	-1	-1	-1	1	1
1	-1	-1	-1	-1	-1	1	-1	1	1	-1	-1	1	-1	1	1	-1	1
1	-1	-1	-1	-1	-1	1	-1	-1	1	1	1	-1	1	-1	-1	1	1
1	-1	-1	1	1	1	-1	1	1	-1	-1	-1	1	-1	-1	1	1	6
1	-1	-1	1	1	1	-1	1	-1	-1	1	1	-1	1	1	-1	-1	4
1	-1	1	-1	1	-1	1	1	1	-1	-1	-1	-1	1	1	-1	-1	2
1	-1	1	-1	1	-1	1	1	-1	-1	1	1	1	-1	-1	1	1	4
1	-1	1	1	-1	1	-1	-1	1	1	-1	-1	-1	1	-1	-1	1	6
1	-1	1	1	-1	1	-1	-1	-1	1	1	1	1	-1	1	1	-1	4
1	1	-1	-1	1	1	-1	-1	-1	-1	-1	1	-1	-1	1	1	1	4
1	1	-1	-1	1	1	-1	-1	1	-1	1	-1	1	1	-1	-1	-1	5
1	1	-1	1	-1	-1	1	1	-1	1	-1	1	-1	-1	-1	1	-1	2
1	1	-1	1	-1	-1	1	1	1	1	1	-1	1	1	1	-1	1	1
1	1	1	-1	-1	1	-1	1	-1	1	-1	1	1	1	1	-1	1	4
1	1	1	-1	-1	1	-1	1	1	1	1	-1	-1	-1	-1	1	-1	5
1	1	1	1	1	-1	1	-1	-1	-1	-1	1	1	1	-1	-1	-1	3
1	1	1	1	1	-1	1	-1	1	-1	1	-1	-1	-1	1	1	1	3

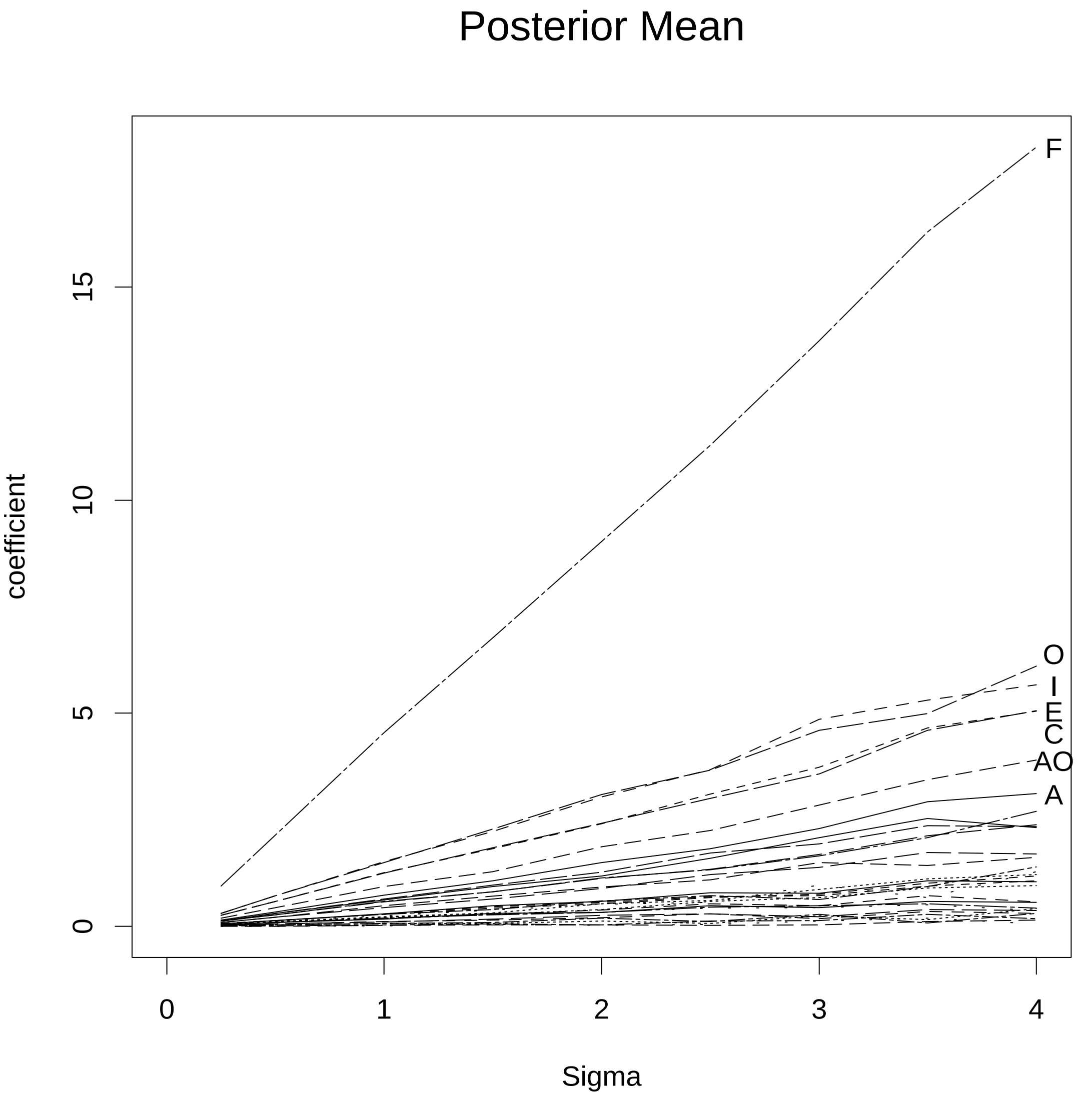
Table 3.6 Model choice criteria for various models for the handle data

junk	df	LR	AIC	BIC	PBF
γ,F	6	70.27	10.27	-33.70	14.34
γ,F,I	7	62.81	4.81	-37.70	11.30
γ,F,O	7	63.13	5.13	-37.38	11.46
γ,F,I,O	8	48.23	-7.77	-48.82	4.70
γ,C,E,F,I,O	10	36.99	-15.01	-53.12	0.47
γ,C,E,F,I,O,AO	11	28.93	-21.07	-57.72	-2.86
γ,A,C,E,F,I,O,AO	12	24.39	-23.61	-58.79	-4.44
γ,A,C,E,F,I,O,P,AO	13	16.37	-29.63	-63.35	-7.76
γ,A,C,E,F,I,K,O,P,AO	14	8.86	-35.14	-67.38	-10.82
γ,A,C,E,F,I,K,O,P,AO,AQ	15	0.37	-41.63	-72.41	-14.37
γ,A,C,E,F,I,K,O,P,AO,AQ,DM	16	0.02	-39.98	-69.30	-13.85
γ,A,C,D,E,F,I,K,O,P,AO,AQ,DM	17	0.00	-37.99	-65.84	-13.17
γ,A,C,D,E,F,I,K,O,P,AD,AO,AQ,DM	18	0.01	-35.99	-62.37	-12.47

available from standard packages, and the diffuse prior method described in Section 2.3.2 was used to obtain the maximum likelihoods. These measures are given in Table 3.6 for several of the models considered. Note that the log-likelihood of the largest model in the table is exactly zero (ignoring sampling variation due to the maximization method). This implies a perfect fit to the data. It is immediately clear that the model A, C, E, F, I, O, AO is not optimal under any of the criteria in this table. The choice of a smaller model is deliberate, since the "best model" (15 parameters) has almost one parameter for every two data points, and predicts the observed outcomes with almost no error. The preference of the criteria for larger models is due in part to the fact that they do not incorporate prior information. If further studies were to be carried out, consideration of larger sets of terms would be justified. Since optimization will be carried out using the model fitted to the same data, a smaller 12 parameter model is chosen. Note that the 11 parameter model was not considered because it contains an AO interaction but no A main effect. Models containing interactions but not corresponding main effects are typically difficult to explain.

One aspect of the model fitting not mentioned in the previous section is robustness to prior variance. Informative priors shrink the posteriors towards zero, and produce finite parameter estimates. The degree of shrinkage corresponds to the prior variance, and since knowledge of prior variance may not be exact, the dependence of the posteriors upon it should be examined. This dependence was mentioned by Hamada and Wu (1992) and may be studied via a new graphical display, in which summaries of all marginal posteriors are plotted as a function of prior variance. For example in Figure 3.3, the posterior mean of each marginal is plotted for prior $\sigma_\beta \in (0.25, 4)$. In this problem, some of the MLE's are infinite, so as the prior becomes more diffuse, the posterior means diverge, approaching

Figure 3.3 Robustness of posteriors to prior variance when MLE's are infinite, injection molding experiment



the MLE's. The magnitude of the summary changes over prior variances, but the rankings of factor effects do not. This robustness to prior information increases confidence in the model selection process, which is carried out at a single value of $\sigma_{\boldsymbol{\beta}}$ (1 here). This type of plot offers many possibilities, including the ability to study the effect of changing prior variance for one set of parameters with the prior variance held constant for the remaining parameters.

The rationale for using an upper limit of $\sigma_{\boldsymbol{\beta}} = 4$ is that $\Pr(\beta_i \in (-8, 8)) = 0.95$ (for the prior). Thus with probability 0.95, the largest possible effect of shifting X_i from -1 to 1 would be a shift in cumulative probability from $\Phi(-8)$ to $\Phi(8)$, or 6×10^{-16} to $1 - 6 \times 10^{-16}$. The wide range of cumulative probabilities implies that $\sigma_{\boldsymbol{\beta}} = 4$ is an adequately large upper bound.

It is the optimization stage of this example that is interesting, since the on-target problem presents challenges not present in the other data sets. In "higher the better" experiments it is often enough to determine the direction of each effect, and then set that factor to a corresponding high or low level, so as to push the probability mass to the high end of the scale. In an on-target problem, things are not that simple. Since the target is a middle category, the setting of each factor will depend on the settings of the others. For example, if one effect is positive, and the other negative, they could be canceled out to get on target. There are two ways to do this (considering settings of only ± 1 levels of factors), and as the number of terms in the model increases, there are many different ways to accomplish similar goals. Thus, for a single on-target goal, there are likely to be several "optimal" factor settings. This is in fact an advantage, since the cheapest settings could be chosen or a secondary criterion used to distinguish between the settings.

These concepts are more clearly illustrated in the optimization stage for this experiment. Since little knowledge of the process is available, consider optimizing over high and low levels (± 1) of the factors. This means that there are $2^6 = 64$ settings to consider. All optimization is relative to the posterior distribution of $\mathbf{p}$, the proportion of parts in each response category. The goal is factor settings that produce on-target posteriors, namely ones with most parts in response group four. While there may be other favorable properties, assume that it is crucial that settings produce a high proportion of on-target parts. Promising settings are identified by considering the expected posterior proportion of parts in response category four. The five most promising settings are displayed in Table 3.7, along with means and standard deviations of the posteriors for $\mathbf{p}$. The first four setting appear quite similar, and the fifth has fewer on-target parts. The structure of the four best settings implies that $F = 1, I = -1, O = 1, C = -E$ must be chosen. Choice of A appears unimportant. If the only goal is to have as many parts on target as possible then the C, E combination may be chosen on a cost basis. If there are other priorities, they may be used to choose between the five settings. Possible secondary criteria might include:

- **Low variability in the proportion of on-target parts.** If one of the optimal settings has less variability in the proportions, that setting would be favored. In this case, there is little distinction between the first four settings, and the fifth has slightly

Table 3.7 Posterior expectations and variances of proportions for five optimal settings, injection molding experiment.

#	Factor Settings						Posterior Means and Standard Deviations					
	A	C	E	F	I	O	1	2	3	4	5	6
45	-1	-1	1	1	-1	1	0 (0.01)	0.02 (0.05)	0.06 (0.07)	0.82 (0.14)	0.10 (0.12)	0 (0)
43	-1	1	-1	1	-1	1	0 (0.01)	0.02 (0.05)	0.06 (0.07)	0.82 (0.14)	0.10 (0.13)	0 (0)
44	1	1	-1	1	-1	1	0 (0.01)	0.03 (0.06)	0.08 (0.09)	0.81 (0.14)	0.07 (0.10)	0 (0)
46	1	-1	1	1	-1	1	0 (0.01)	0.04 (0.06)	0.08 (0.09)	0.81 (0.14)	0.07 (0.10)	0 (0)
58	1	-1	-1	1	1	1	0 (0.00)	0.01 (0.04)	0.04 (0.07)	0.77 (0.18)	0.17 (0.19)	0 (0)

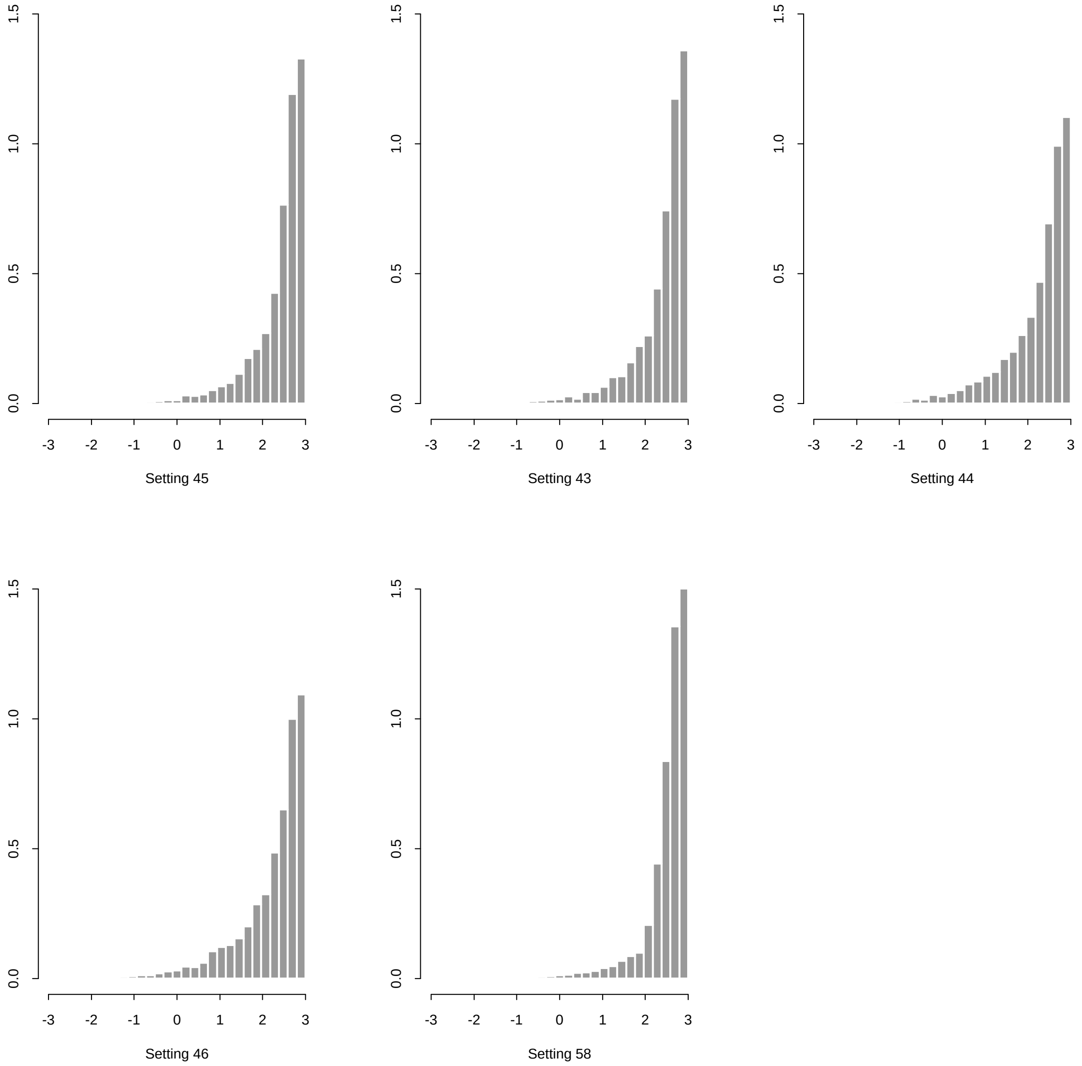
more variation.

- **Other outcome groups.** Slight underfills or overfills (categories 3 and 5) may have a smaller loss associated with them than more extreme outcomes. Hence the proportions of parts in all six categories should be examined, either by numeric methods or plots (not shown here). Table 3.7 indicates that settings 43-46 have 2% to 4% mass outside categories 3,4,5 while setting 58 has only 1%. If parts outside this range must be scrapped (at great cost), these four settings would be less preferable.

- **A specific criterion based on projections.** The six dimensional distribution of proportions may be examined via projections of various kinds. Examples already used include proportions in a single category (like the projection (0,0,0,1,0,0)) or in two categories (Figure 3.2 in the foam example). Projections of the posterior proportions onto other more complicated (but meaningful) subspaces may be used to distinguish different settings. The meaning of the projection can simply be an informative direction (for example (0,0,1,2,1,0), which represents a linear combination measuring the concentration in and near category four), or it can have a cost interpretation. If each element of the projection vector is a profit associated with that category, then the projection is a posterior distribution for earnings. This follows, since the projection is a sum of the product of proportions and profits for each outcome. So for example, parts that were underfilled might have to be scrapped (a loss) while parts above target could be reworked (a smaller profit than on-target). The projection (-3,-3,-3,3,2,1) could represent associated earnings, and the distribution of profits used to evaluate the five settings. Note that this projection translates an "on target" problem to a "larger the better" problem, since profit is to be maximized. The profit distribution is shown in Figure 3.4. Setting 58, which has a smaller on target proportion, has the best profit distribution, since the tail at the lower profit end trails off faster than the other four examples. Note that this specific scoring is chosen as an illustration.

The reader is likely to notice similarities between the projections introduced here and the scored ANOVA approach, since in both cases scores are associated with each outcome

Figure 3.4 Projected posterior proportions for projection (-3,-3,-3,3,2,1) = scrapped underfills and reworkable overfills, injection molding experiment



category. The distinction is that projections are used as an interpretive device, and are not central to the model fitting process. With scored ANOVA, the scores influence both the model selection and optimization stages, whereas in this approach, the scores are added after a model has been identified. This philosophy is similar to the response modeling philosophy, which advocates fitting a model to the original data and not summaries. In robust design, the summaries might be signal to noise ratios or log sample variances, and the data the original response; here the summaries are scores, and the category counts the original data. Only after the model is fit are interpretive summaries used (in response modeling, mean and variance, here projections and other interpretive tools).

The overall best one or two settings will not be chosen from these criteria, because of the lack of knowledge about the real goals of the experiment, other than to produce on-target parts. In conclusion, note that for the on-target problem, it is likely that several settings will produce comparable optimal results. This may be seen as an advantage, since if there are no secondary goals the cheapest settings may be used; otherwise secondary goals may be used to identify the best factor level settings.

Chapter 4

Variable Selection Introduction

4.1 Introduction

It was shown in the previous chapters that variable selection is a central issue to fitting models to ordinal data. The case studies of Chapter 3 illustrate a number of shortcomings of the variable selection approach of Section 2.3.1, including:

1. **Additional prior information unused.** In the two examples, the models suggested by model choice criteria were sometimes in conflict with prior beliefs about what constitutes a "reasonable" model. For example, in the injection molding data (Section 3.2) all criteria favored models with many predictors, which seems unreasonable in light of the small sample size. While this is partly due to the criteria used, the priors have no explicit way to incorporate information about the number of terms that are active.

2. **Relations between predictors.** In both examples, some of the model terms (e.g. interactions) were constructed from the basic variables. It is a standard convention that if such composite variables are important, some or all of the variables that compose them are included. For example, in the handle data, the importance of an AO interaction meant that only models which include terms A and O were considered. This restriction entered into the modeling process as an informal belief, rather than a component of the priors.

3. **Cumbersome and incomplete model search.** When the MLE exists, automatic (albeit non-Bayesian) stepwise procedures can be used to identify promising models. These procedures consider only a small subset of the entire model space, leaving the possibility that a sub-optimal model will be found. Furthermore, packages utilizing these procedures fail altogether when the MLE does not exist. Manual or semi-automatic searches based on the Bayesian model of Section 2.2.2 are possible but computationally expensive. Unlike ANOVA, which can marginally identify variables as important, variable selection for GLMs must proceed by fitting each candidate

model. The reason for this is that the nonlinear link function introduces correlations in the posteriors of the regression coefficients, which implies that dropping a term from the model will change the other coefficients. What is needed is an automatic Bayesian model search technique that considers the entire model space in a reasonable amount of time.

4. **Approximate choice criteria.** The model choice criteria used are based on asymptotic results because exact calculations were not possible. The approximations used do not incorporate prior information in any explicit way, and are probably quite inappropriate for the small samples encountered in industrial data. A desirable feature for a new procedure would be exact (or more accurate) model choice criteria that incorporate prior information.

The purpose of the second part of the thesis is to develop an automatic Bayesian variable selection algorithm which addresses these problems. In this chapter, a automated technique that will form the basis for several generalizations is introduced. An un-modified version of this algorithm addresses points three and four, and point one in part. The generalization considered in the next chapter will address point two (a point with considerable breadth of application), and more fully address point one. Before details are given, some general discussion of the variable selection problem is given.

In regression models with numerous predictors, the issue of variable selection arises naturally. Given a set of p predictor variables, the basic goal is to identify a subset which predict the response well. Quite often, automatic techniques are used when p is large, since there are a total of 2^p possible models from which to choose. The majority of these techniques involve comparison of models based on some type of goodness-of-fit statistic such as R^2, AIC, C_p, etc. When the number of predictors (and hence the number of candidate models) is small enough, an exhaustive search may be carried out. This approach can become computationally very expensive for a large number of predictors, and algorithms that consider a subset of all possible models may instead be used. For example, stepwise selection techniques compare the current model to those which are close to it, typically those with one term removed or one term added. By successively adding and removing terms, it is hoped that a good model will be found. Miller (1991) reviews model selection methods, including all subsets and stepwise methods.

Typically, all 2^p models are either considered, or could be considered by these automated techniques. Quite often some of these models are not likely to be used, even if they provide the best fit for the data. For example, some statisticians adopt the convention that a model containing an interaction should also contain the corresponding main effects. A common justification of this convention is that the models are easier to interpret. It is common practice to "clean up" the best model by adding in main effects corresponding to interactions in the model, or remove terms which are too complicated or difficult to explain. Not only is this technique time consuming, but it is done after the fact, rather than as an integrated part of the model selection procedure.

In doing so, statisticians are informally enforcing beliefs and preconceptions about which subsets of variables are likely to provide a good model. These beliefs implicitly specify relationships between predictors which are not usually recognized by automatic selection procedures. Possible reasons why these constraints are not built into subset selection algorithms include the additional bookkeeping necessary, and the inflexibility of the resulting procedures. That is, there is not enough latitude allowed by simply disallowing certain models because they seem unlikely. The classification of models into possible and impossible groups is too coarse, and one may be reluctant to consign models to the latter class unless it is quite clear that they are not conceivable.

This chapter proposes a richer, more mathematical way of expressing beliefs about the relationships between predictor variables. Instead of classifying a model as either "allowed" or "forbidden", a Bayesian approach is used to assign degrees of belief to each model in a structured way that accounts for relationships between predictors.

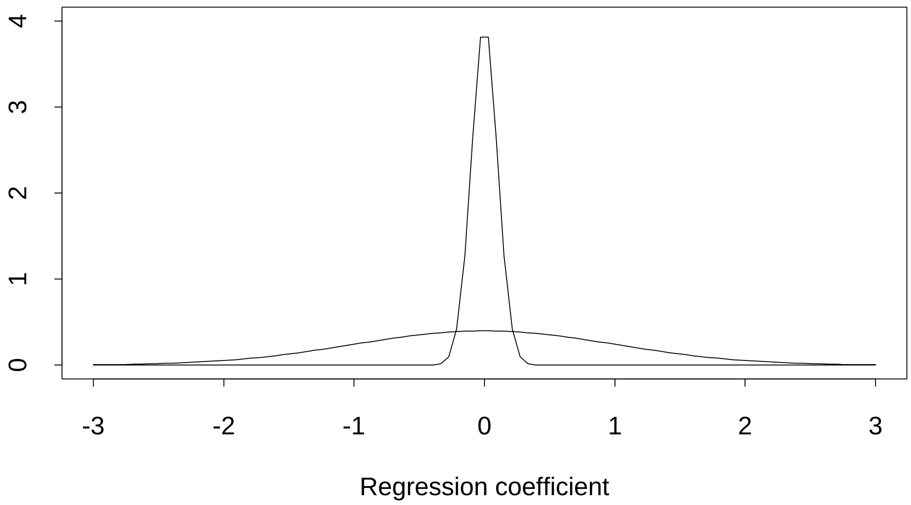
These generalizations deal with several different types of common dependencies between model terms. The first relates to two-way (and more general) interactions, and expresses the probability that an interaction will be active conditionally on whether its parents are active. A second type of relation groups like terms together and considers only models with all or none of these terms active. Relations between terms which are mutually exclusive are developed, and applied to the problem of selecting transformations for predictors, and mutually exclusive hypotheses. Finally, global constraints on the model are introduced as a method of re-weighting models based on "larger scale" properties, such as the total number of active terms. Since the methodology developed here is applicable to a broader class of problems than just ordinal data (including linear regression and GLMs), the language used here will refer to any of these situations.

Although the probability structures are general enough to be used for different Bayesian model selection algorithms, the theory is developed with a certain technique in mind. The approach of George and McCulloch (1993), henceforth denoted by GM, parameterizes the model selection problem by associating latent Bernoulli variables $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)$ with each predictor. These latent variables determine whether a factor is *active* or *inactive*, and the priors placed on $\boldsymbol{\delta}$ represent beliefs about the relationships between predictors. It is via these priors that preferences (as opposed to acceptable/unacceptable classifications) for certain models will be expressed. An additional gain of using the approach of GM is that it carries out a Bayesian equivalent of all subsets selection, even in situations where the model space is quite large.

4.2 A Current Technique

In this part of the thesis an extension of a numeric technique for calculation of $\Pr(M_i|Y)$ is considered. The details of this approach are given in GM. It will be assumed that the criterion of interest is the posterior probability of a model conditional on the data, although the other expressions could be calculated using the results from the numeric procedure. The

Figure 4.1 Mixture prior for β_i



idea of calculating the posterior probability of models has several precursors, including the work of Box and Meyer (1986, 1993) and Mitchell and Beauchamp (1988).

The approach in GM can be outlined as follows for the simplest case of linear regression with normal errors,

$$Y = X'\boldsymbol{\beta} + \epsilon. \tag{4.1}$$

The central concept is to introduce an unobserved vector $\boldsymbol{\delta}$ of zeros and ones of the same length as $\boldsymbol{\beta}$. The components of this vector represent the importance of the corresponding regressor variables. That is, if $\delta_i = 0$, then the magnitude of β_i is small, and the corresponding predictor is "inactive". If $\delta_i = 1$, then the magnitude of β_i is large, and the predictor is "active". Mathematically, this is accomplished by defining a mixture prior for the coefficients $\boldsymbol{\beta}$:

$$f(\beta_i|\delta_i) = \begin{cases} N(0, \tau_i^2) & \text{if } \delta_i = 0 \\ N(0, (c_i\tau_i)^2) & \text{if } \delta_i = 1 \end{cases} \tag{4.2}$$

When $\delta_i = 0$, β_i is tightly centered around 0, and will not have a large effect. The much larger variance ($c_i >> 1$) allows the possibility of a variable having a large influence. Figure 4.1 presents the case $\tau_i = 0.1, c_i = 10$. The parameters c_i and τ_i must be chosen to represent a "small" effect, and how many times larger a "large" effect would be. The choice of these parameters is obviously closely related to the location and scale of the predictor variables. GM propose automatic methods for the choice of these parameters, and also recommend the use of several different choices, to see how differing choices influence the conclusions made.

The vector $\boldsymbol{\delta}$ may be thought of as containing information relevant to the importance of the predictors. Thus the posterior distribution for $\boldsymbol{\delta}$ will indicate which models are suggested by the data and prior knowledge. Note that henceforth the term model will refer to any subset of predictors under a fixed structural model (such as normal linear regression).

This specific parameterization is chosen so that a Gibbs sampling approach may be used to obtain the posterior for $\boldsymbol{\delta}$. In GM, the conditional distributions

$$f(\boldsymbol{\beta}|\boldsymbol{\delta},\boldsymbol{\phi},Y),\, f(\boldsymbol{\phi}|\boldsymbol{\beta},\boldsymbol{\delta},Y),\text{ and } f(\boldsymbol{\delta}|\boldsymbol{\beta},\boldsymbol{\phi},Y) \tag{4.3}$$

are used to obtain samples from the posterior distribution. Here, the parameter $\boldsymbol{\phi}$ represents all other parameters, such as the variance of ϵ in the linear model case, or for ordinal models,

$$\boldsymbol{\phi} = \{\mathbf{Z},\gamma\}.$$

Sampling $\boldsymbol{\delta}$ in (4.3) is replaced by sub-steps in which each element δ_i is sampled conditional on the remaining elements $\boldsymbol{\delta}_{(-i)}$. While this may slow the convergence of the algorithm, it greatly simplifies the task of sampling from conditional distributions. The draw for $\boldsymbol{\beta}$ is a multivariate normal, and for each individual δ_i it is a Bernoulli draw. The drawing of $\boldsymbol{\phi}$ depends on the specific problem, and may be subdivided into several conditional draws (for example $\mathbf{Z}|\mathbf{Y},\gamma,\boldsymbol{\beta}$ and $\gamma|\mathbf{Y},\mathbf{Z}$ in the ordinal case). Each draw depends on the most recent values of the other parameters and the data. Further details of this procedure may be found in GM, and in George, McCulloch, and Tsay (1994) in the case of binary probit regression. Gibbs algorithm details for the generalization of GM are given in Section 5.6.2.

What is of interest here is the choice of prior for the vector $\boldsymbol{\delta}$. Although their theory is general, in practice GM suggest the use of an independence prior:

$$f(\boldsymbol{\delta}) = \prod_{i=1}^{p} p_i^{\delta_i}(1-p_i)^{1-\delta_i}, \tag{4.4}$$

where $p_i = \Pr(\delta_i = 1)$. That is, the importance of any variable is independent of the importance of any other variable. This is a very parsimonious representation of prior knowledge, and in many situations it is quite accurate and appropriate.

There are some cases in which this independence prior is not appropriate, even when there is not a large amount of prior knowledge. For example, consider the predictors A, B, and their interaction AB as candidate terms. If it was known that neither of the main effects A or B was important, one might suspect that AB would also not be important, even without evidence from data. Clearly, the independence prior for $\boldsymbol{\delta}$ is no longer valid, and a richer and more flexible prior is needed to represent relations among elements of $\boldsymbol{\delta}$. The next chapter of the thesis develops priors to represent these and other relations between predictor terms.

Chapter 5

Priors for Related Predictors

In this chapter, priors that incorporate relations between predictors are developed. These relations are specifically for $\boldsymbol{\delta}$, since all the relations discussed will be related to model selection. Two kinds of assumptions will be represented in these priors - structural assumptions (such as independence) and probabilistic ones. Both assumptions may be viewed as qualitative representations of commonly utilized principles of variable selection. In addition to the simple case of two way interactions given in the previous section, relations for general polynomial interactions, grouped variables, competing variables, and restrictions on the number of active terms will be considered.

5.1 Relations for Two-Way Interactions

In this section, a simple example of non-independent priors will be given for a case in which there are three main effects A, B, C and three two-way interactions AB, AC, and BC. The goal is to formulate a prior for the importance of these variables in a fashion that represents relations among them.

The joint density of $\boldsymbol{\delta}$ may be factored as follows:

$$\begin{aligned}\Pr(\boldsymbol{\delta}) &= \Pr(\delta_A, \delta_B, \delta_C, \delta_{AB}, \delta_{AC}, \delta_{BC}) \\ &= \Pr(\delta_A, \delta_B, \delta_C)\Pr(\delta_{AB}, \delta_{AC}, \delta_{BC}|\delta_A, \delta_B, \delta_C)\end{aligned} \tag{5.1}$$

This distribution can be simplified by assuming that $\delta_A, \delta_B, \delta_C$ are independent, and that conditional on $(\delta_A, \delta_B, \delta_C)$, the interactions $(\delta_{AB}, \delta_{AC}, \delta_{BC})$ are independent:

$$\begin{aligned}\Pr(\boldsymbol{\delta}) = \Pr(\delta_A)&\Pr(\delta_B)\Pr(\delta_C) \times \\ &\Pr(\delta_{AB}|\delta_A, \delta_B, \delta_C)\Pr(\delta_{AC}|\delta_A, \delta_B, \delta_C)\Pr(\delta_{BC}|\delta_A, \delta_B, \delta_C)\end{aligned} \tag{5.2}$$

This expression can be further simplified by assuming that the importance of an interaction depends only on the importance of those main effects from which it was formed. That is, δ_{AB} depends on δ_A and δ_B, but not δ_C. This simplifies the last three terms in (5.2) to

$$\Pr(\boldsymbol{\delta}) = \Pr(\delta_A)\Pr(\delta_B)\Pr(\delta_C)\Pr(\delta_{AB}|\delta_A, \delta_B)\Pr(\delta_{AC}|\delta_A, \delta_C)\Pr(\delta_{BC}|\delta_B, \delta_C) \tag{5.3}$$

These two assumptions simplify the structure of the joint density for $\boldsymbol{\delta}$ greatly. Their importance is not simply a matter of mathematical convenience; these assumptions correspond to principles for variable selection that are commonly utilized. Both principles can be related to the hierarchical nature of the different model terms. Namely, the main effects are of the lowest (or simplest) order, and two-way interactions are of a higher (and thus more complex) order. The first assumption is that terms of a given order are independent, conditional on all terms of a lower order. This will be called the *conditional independence principle.* The second assumption comes from a notion of inheritance between terms of different orders, and states that a higher order term does not depend on all lower order terms, but only those from which it inherits. This will be called the *inheritance principle.*

The exact nature of this dependence on "parent" terms is defined by the components of the joint probability in (5.3). For example, the probability that the term AB is active $\Pr(\delta_{AB} = 1|\delta_A, \delta_B)$ may take on four different values, depending on the values of the pair (δ_A, δ_B):

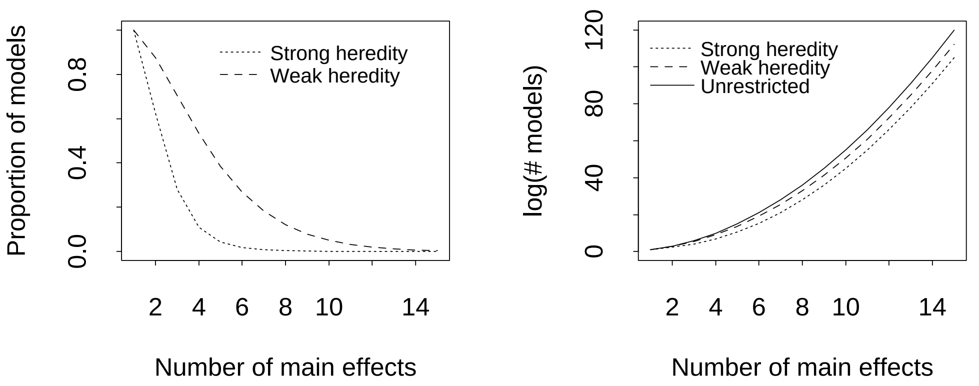
$$P(\delta_{AB} = 1|\delta_A, \delta_B) = \begin{cases} p_{00} & \text{if } (\delta_A, \delta_B) = (0,0) \\ p_{01} & \text{if } (\delta_A, \delta_B) = (0,1) \\ p_{10} & \text{if } (\delta_A, \delta_B) = (1,0) \\ p_{11} & \text{if } (\delta_A, \delta_B) = (1,1) \end{cases} \tag{5.4}$$

The choice of these values may be made to represent different principles of variables selection. For example one might choose $(p_{00}, p_{01}, p_{10}, p_{11}) = (0, 0, 0, p)$. This would correspond to the prior belief that for an interaction to be active, all (in this case both) corresponding main effects must also be active. This principle has several names, including the "marginality assumption" (McCullagh and Nelder, 1989), and "well formulated models" (Peixoto, 1987, 1990). A less restrictive choice of the conditional probabilities might be $(p_{00}, p_{01}, p_{10}, p_{11}) = (0, p_1, p_2, p_3)$, corresponding to the belief that an interaction may enter the model if one or more corresponding main effect is in the model. Hamada and Wu (1992) call this the "effect heredity" principle. A new naming convention will be adopted here, with the *strong heredity principle* corresponding to $(0, 0, 0, p)$ and the *weak heredity principle* corresponding to $(0, p_1, p_2, p_3)$.

In both cases, it seems that there is likely to be an implicit ordering in the probabilities, with $p_{00} \leq (p_{01}, p_{10}) \leq p_{11}$. It should also be noted that when there is little prior knowledge the same four probabilities could be used for all interaction terms. On the other hand, there may be prior knowledge that is specific to certain variables or collections of variables, which could also be used in (5.3). Some specific examples, including non-interacting subsystems, are given in Chapter 6.

It has already been mentioned that different choices of prior probabilities such as p_{11} represent heredity principles. The magnitude of the values chosen may be related to another variable selection principle, namely "effect sparsity", a term coined by Box and Meyer (1986). By making the prior probability of an active predictor small, the principle of the "magnificent few and the trivial many" predictors is expressed. This principle will be called

Figure 5.1 Number of models for restrictions on interactions given parents



the *sparsity principle*. Note that it may be applied to both main effects and higher order terms.

It is clear that by setting one or several of p_{00}, p_{01}, p_{10} to be exactly zero, the model space is reduced. The number of models with positive mass can be calculated for varying rules; here calculations for strong and weak heredity principles are presented. Suppose there are m main effects, and all possible two-way interactions are considered in the model. Then there are $m + \binom{m}{2}$ candidate terms, making a total of $2^{m+\binom{m}{2}}$ models. If strong heredity is assumed, (i.e. $p_{00} = p_{01} = p_{10} = 0$), then it can be shown that there are

$$\sum_{i=0}^{m} \binom{m}{i} 2^{\binom{i}{2}} \tag{5.5}$$

models with nonzero mass. In the more relaxed case of weak heredity, (i.e. $p_{00} = 0$, others > 0), there are

$$\sum_{i=0}^{m} \binom{m}{i} 2^{mi-i(i+1)/2} \tag{5.6}$$

models with nonzero mass. Proofs of (5.5) and (5.6) are given at the end of this section, on page 44.

To give a sense of what this means, the total number of models is compared to the number of models under these two assumptions in Figure 5.1. Two representations of the size of the model space are given in this figure. In the first, the proportion of the model space with nonzero prior probability is plotted. Clearly as the number of main effects increases, a smaller percentage of the whole model space is allowed, and the more restrictive strong heredity principle (the finely dotted line) reduces the proportion most rapidly. In the second plot, the log of the number of models is plotted against the number of main effects. A base two log is used so that the vertical scale is an "effective number of predictors". That is, with fifteen main effects, there are 120 terms, and the same number of models as if there were effectively 112 terms under weak heredity and 105 under strong heredity. This is informative, since it indicates that although the proportion of allowed models decreases with the number of main effects, there are still a very large number of models in all three cases. Other calculations could be presented for more complicated relations or for subsets of all possible interactions, but this gives a flavor for the gains of these restrictions.

A more interesting issue is the concentration of prior mass among models. There could be many models which are a priori possible, but very unlikely. In such a situation, the number of possible models might be close to the unrestricted case, but prior mass is focused on a much smaller subset of the model space. A further discussion of these issues, along with graphic and numeric methods for measuring the concentration of prior mass are discussed in Chapter 7.

It should be noted that sometimes a higher order term (such as an interaction) will be active without a lower order term (such as a main effect). For example, in Newtonian physics, there is the well-known relation $F = MA$, relating the force exerted on an object to its mass and acceleration. Another example is the equation $E = \frac{1}{2}MV^2$ relating kinetic energy to mass and velocity. In both cases, there is an interaction (and in the latter case, a higher order polynomial term - see the next section) with no main effects. These examples may seem simplistic, but that is in part because these relations are already known - presumably in the search for (as yet) unknown relations, it is conceivable that strong or weak heredity will not hold.

On the other hand, hierarchical structure will often be appropriate, especially if the model is used as an approximation. In that situation, first order terms represent linear approximations, and the higher the order of the term, the more complicated and nonlinear the model for the response will be. The presence of higher order terms will typically mean more variation in prediction (especially interpolation or extrapolation to unexplored areas), and so it seems natural to include higher order terms only after lower order terms have been given every chance to model the response.

The point is that while a hierarchical ordering is often sensible, it is not always. A compromise may be struck between the commonly used hierarchical structure, and the (perhaps small) possibility of interactions without parents. Instead of setting any of the probabilities $p_{00}, p_{01}, p_{10}, p_{11}$ to be exactly zero, some small number ε could instead be used. Thus, strong heredity $(0, 0, 0, p_{11})$ would be relaxed to $(0, \varepsilon, \varepsilon, p_{11})$ or $(\varepsilon, \varepsilon, \varepsilon, p_{11})$, and weak heredity $(0, p_{01}, p_{10}, p_{11})$ would be relaxed to $(\varepsilon, p_{01}, p_{10}, p_{11})$. When the data present overwhelming evidence contrary to heredity, the posteriors for $\boldsymbol{\delta}$ will reflect this (as would be the case for $F = MA$). However, if the evidence is not overwhelming, the weak or strong heredity principles would outweigh the evidence. Note that although the terms strong and weak heredity strictly correspond to $\varepsilon = 0$, they will still be used with small nonzero values of ε, since most prior mass is still concentrated on models allowed by weak or strong heredity principles.

Proof of combinatorics for (5.5) and (5.6)

In both proofs, assume that there are m main effects, each of which may be active or inactive. Suppose also that each of the $\binom{m}{2}$ two-way interactions between these effects are considered as possible predictors. For a set of active main effects, strong and weak heredity will require that certain interactions must be inactive, while others could be either active or inactive. This latter set will be referred to as interactions that are *eligible* to be active.

Strong heredity (5.5):

The strong heredity principle states that an interaction is eligible to be active only when both main effect parents are active. Then for a set of i active main effects there are $\binom{i}{2}$ interactions eligible to be active. Since each may be active or inactive independently of the others, there are a total of

$$2^{\binom{i}{2}}$$

distinct models for each different set of i active main effects. There are $\binom{m}{i}$ different sets of i active main effects, so summing i from 0 to m yields a total of

$$\sum_{i=0}^{m} \binom{m}{i} 2^{\binom{i}{2}} \tag{5.5}$$

distinct models under strong heredity.

Weak heredity (5.6):

The only difference from the previous proof is the number of eligible two-way interactions for each set of i active main effects. The weak heredity principle states that that a two-way interaction is eligible to be active if one or more of the main effect parents are active. Consider a set of i active main effects labeled $m_1, m_2, \ldots, m_i$. All interactions between each of $m_1, \ldots, m_i$ and the other $m-1$ main effects are eligible - the trick is to count each only once. Starting with m_1, there are $m-1$ eligible two-way interactions. Then m_2 will contribute an additional $m-2$ uncounted eligible two-way interactions (skipping the already counted $m_1 m_2$ interaction) , and so on up to $m-i$ new interactions for m_i. That is, there will be a total of

$$\sum_{j=1}^{i} (m-i) = mi - \frac{i(i+1)}{2}$$

interactions eligible to be active for each set of i active main effects. This means that there are

$$2^{mi-i(i+1)/2}$$

different models for each distinct set of i active main effects. There are $\binom{m}{i}$ such distinct sets of i active main effects, so summing over i yields

$$\sum_{i=0}^{m} \binom{m}{i} 2^{mi-i(i+1)/2} \tag{5.6}$$

distinct models with nonzero mass.

5.2 Hierarchical Polynomial Interactions

The principles and methods for expressing priors used in the previous section may be generalized to the case of models containing interactions between an arbitrary number of terms, each of an arbitrary order. As in the previous section, a simple example will be given, and then the principles and structure presented in a general form.

Consider a model with the terms $A^2B^2, AB^2, A^2B, A^2, AB, B^2, A$, and B. In order to write out the prior for $\boldsymbol{\delta}$, hierarchical and inheritance relations must be defined among the terms considered. One possible ordering can be given by the total exponent of all components of each term. Thus AB^2 and A^2B are of the same order, and of a lower order than A^2B^2. A term is said to *inherit* from a collection of lower order terms if it is equal to the product of those lower order terms. Thus, A^2B inherits from A^2, AB, A, B but not B^2. These relations can be expressed using a directed graph, as in Figure 5.2. The principles of conditional independence and inheritance may be used here to simplify the prior on $\boldsymbol{\delta}$. In this case, the density may be factored into a chain of terms of a given order, each conditional on all terms of a lower order. The conditional independence principle states that all terms of a given order are independent, conditional on all lower order terms. The inheritance principle is then used to state that a term depends only on those terms from which it inherits.

A third principle may achieve further simplification. A term *inherits immediately* from another term if it inherits from that term, and the term is of the next lowest order. That

Figure 5.2 Ordering and inheritance relations among polynomial interactions

$$\begin{tikzcd}[column sep=small]
 & & A^2B^2 & & \\
 & AB^2 \arrow[ur] & & A^2B \arrow[ul] & \\
B^2 \arrow[ur] & & AB \arrow[ul] \arrow[ur] & & A^2 \arrow[ul] \\
 & B \arrow[ul] \arrow[ur] & & A \arrow[ul] \arrow[ur] &
\end{tikzcd}$$

is AB^2 inherits immediately from AB and B^2, and it inherits (but not immediately) from A and B. The terms "parent" and "child" will be used in the context of immediate inheritance. The *immediate inheritance principle* is then defined as the assumption that given the importance of its parents, the importance of a child term is independent of all other terms.

The implication of this third principle is that the importance of a given term depends on relatively few terms (in comparison to inheritance, which would have high order terms that depend on many lower order terms). In the current example, the importance of each term depends on either 0, 1, or 2 other parents. The choice of prior probabilities can again reflect beliefs such as sparsity and weak or strong heredity. Strong heredity would now require all parents to be active for a child to be active; weak heredity would require only one active parent.

The obvious justification of allowing a child in the model with only one active parent is that the model is more flexible. On the other hand, models containing all parents of a child are easier to interpret, and have a nice property of invariance to linear transformations of the predictors. For example, consider the variables A, B, and AB, and suppose a model model with only the AB interaction is selected. If the transformed variables C=A+1, D=(B+1)/3 were created, and a good model selected, the CD interaction might be again identified as the only important variable. However, this term can be expanded in terms of the original variables A and B, resulting in a model with components of both A, B, and AB. Restricting attention to only those models that contain all parents of a child yields models invariant to linear transformations. Further discussion of the invariance issue and a proof of this property are given in Peixoto (1990).

In situations where there is some natural parameterization for the predictors and reparameterization makes less sense, more justification for weak heredity can be given. For example, if all the predictors are measured in the same units, reparameterization is unlikely, and it is not important for a procedure to be parameterization invariant. An industrial example might be a situation in which some predictors are restricted to a certain region. Reparameterization of this region to the $(-1, 1)$ interval would seem a natural scale, and any other reparameterizations are unlikely to be informative.

Two interesting research topics related to these hierarchical structures are probability networks and image restoration. In the case of the former, directed graphs are used to represent probability networks and the way in which information is propagated across them. While the motivation of this work is different, the networks look similar (see Lauritzen and Spiegelhalter, 1988). The form of the constructed densities is also similar - conditioning and independence assumptions feature prominently. The main similarity is the concept of local dependence - namely the assumption that a node (here a variable like δ_A) has several immediate neighbors which influence its probabilistic behavior, but that its distribution does not depend on more distant neighbors. This principle is also central in image restoration (see Geman and Geman, 1984), where local structure is applied to pixels of an image. In this case, the value of a pixel is assumed to depend only on neighboring pixels.

The power of such an assumption is that it greatly simplifies the model, while still allowing a network of indirect links across all nodes δ_i. The number of parents and children of a given node δ_i is greatly reduced by the immediate inheritance assumption, which may be viewed as a type of local structure. This local structure simplifies the computations necessary to implement this approach. In order to sample from the conditional distribution $\delta_i|\boldsymbol{\beta}, \boldsymbol{\delta}_{(-i)}$, all terms in the prior for $\boldsymbol{\delta}$ not involving the term δ_i are dropped. With the prior structure developed here, all terms are dropped except the probability of the variable itself and expressions involving its children.

The form of the prior on $\boldsymbol{\delta}$ and the simplifications due to principles may be expressed in general as follows. Suppose that the highest order term is of order k. Let $\mathcal{O}_i$ represent all terms of a given order i, $\mathcal{F}(\delta_i)$ represent the family of all lower order terms from which δ_i inherits. Let $\mathcal{P}(\delta_i)$ represent the parents of the term δ_i, namely those terms from which δ_i immediately inherits. Assume that there are p components of $\boldsymbol{\delta}$. Then the density for $\boldsymbol{\delta}$ may be factored by the order of its components:

$$\Pr(\boldsymbol{\delta}) = \prod_{i=1}^{k} \Pr(\mathcal{O}_i|\mathcal{O}_j, j = 1, ..., i-1). \tag{5.7}$$

The conditional independence principle allows the densities for $\mathcal{O}_i$ to be broken into densities for individual elements $\delta' \in \mathcal{O}_i$,

$$\Pr(\boldsymbol{\delta}) = \prod_{i=1}^{k} \prod_{\delta' \in \mathcal{O}_i} \Pr(\delta'|\mathcal{O}_j, j = 1, ..., i-1). \tag{5.8}$$

The inheritance principle reduces the set of terms upon which $\delta' \in \mathcal{O}_i$ depends from $\{\mathcal{O}_j, j = 1, ..., i-1\}$ to the family of the term δ', denoted by $\mathcal{F}(\delta_i)$.

$$\Pr(\boldsymbol{\delta}) = \prod_{i=1}^{k} \prod_{\delta' \in \mathcal{O}_i} \Pr(\delta'|\mathcal{F}(\delta')) = \prod_{i=1}^{p} \Pr(\delta_i|\mathcal{F}(\delta_i)). \tag{5.9}$$

Finally, the immediate inheritance principle reduces the family $\mathcal{F}(\delta_i)$ to the set of parents $\mathcal{P}(\delta_i)$, giving

$$\Pr(\boldsymbol{\delta}) = \prod_{i=1}^{p} \Pr(\delta_i|\mathcal{P}(\delta_i)) = \prod_{i=1}^{p} p_i^{\delta_i}(1-p_i)^{(1-\delta_i)}, \tag{5.10}$$

where in the final expression, $p_i = \Pr(\delta_i = 1|\mathcal{P}(\delta_i))$. This expression has structure similar to the independence prior originally considered in (4.4), with the crucial difference that the terms p_i now are conditional probabilities.

While these relations seem like a sensible structure, it is by no means the only way in which to formulate the prior density. For example, the immediate inheritance principle might not be appropriate, and instead it could be assumed that δ_i depends on its entire family $\mathcal{F}(\delta_i)$. Alternatively, the inheritance relations could be modified so that higher order terms depended on only main effect terms rather than terms of the next lowest order. In the absence of any special knowledge of relations between predictors, the relations given above provide a reasonable structure.

Figure 5.3 The many-to-one mapping of regression coefficients to latent variables

$$
\begin{tikzcd}
\beta_A \arrow[d] & \beta_B \arrow[d] & \beta_C \arrow[d] & \cdots & \beta_{Q_1} \arrow[dr] & \beta_{Q_2} \arrow[d] & \cdots & \beta_{Q_l} \arrow[dll] \\
\delta_A & \delta_B & \delta_C & \cdots & & \delta_Q & &
\end{tikzcd}
$$

5.3 Grouped Predictors

A type of predictor that commonly occurs in regression situations is a variable with many qualitative levels, such as treatment, supplier, or location. Since there is no ordered continuous scale for a multilevel qualitative factor, it is often necessary to create l dummy variables (in the case of a predictor with $l+1$ distinct levels). Quite often (in ANOVA, for example), interest is focused on the primary question of whether these variables make a difference as a group - that is, should the collection of dummy variables be included in or excluded from the model? The other consideration is whether specific terms are important, but by first considering the more general question, the size of the model space is reduced.

Suppose for the purpose of illustration, that there is a group of variables $Q_1, Q_2, ..., Q_l$. The *effect grouping principle* reduces the set of possible values for $\boldsymbol{\delta}$. That is, only

$$(\delta_{Q_1}, \delta_{Q_2}, ..., \delta_{Q_l}) \in \{(0,0,...,0),(1,1,...,1)\}$$

are considered. A simpler way to represent this restriction is to note that there is really only a single scalar δ_Q which determines the importance of $Q_1, Q_2, ..., Q_l$. That is, there is a many-to-one mapping from the vector of regression coefficients $\boldsymbol{\beta}$ to the vector of latent variables $\boldsymbol{\delta}$, as shown in Figure 5.3. The principle of effect grouping need not be constrained to qualitative predictors. There may be other situations in which certain predictors belong to a larger class, and interest focuses on whether that class as a group is influential.

It is interesting to note that the hierarchical relations defined in the previous two subsections can be used in conjunction with the effect grouping principle. The hierarchical relations are defined in terms of the $\boldsymbol{\delta}$s, so δ_Q represents just another variable. For example Q might represent four different suppliers, and A might be temperature. A temperature by supplier interaction would be important if parts from some suppliers were more sensitive to temperature. The interaction of Q with A will have l elements, $AQ_1, \ldots, AQ_l$. Since attention is restricted to models with all or none of the Q terms active, it seems sensible to make the same requirement for $AQ_1, \ldots, AQ_l$. This requirement means that either all or none of the supplier effects change with temperature. It is achieved naturally with the many to one relation in Figure 5.3. Both δ_Q and δ_A are single elements of $\boldsymbol{\delta}$, so the corresponding term for the interaction also has a single element, δ_{AQ}, and this element specifies whether all the terms $AQ_1, \ldots, AQ_l$ are in the model. The only difference from the hierarchical polynomials is that a quadratic term has no meaning for variables like Q. Thus, δ_Q will have no children other than interactions with other variables.

Suppose that $\Pr(\delta_Q = 1|Y)$ is large. The first question that comes to mind is whether only certain components of Q are influential. One obvious approach might be to refit the model using a prior with l independent elements of $\boldsymbol{\delta}$ that correspond to Q. This requires not only a modification of prior knowledge, but additional numeric computation. Instead, there is information in the posterior of the regression coefficients $\beta_{Q_1}, \beta_{Q_2}, ..., \beta_{Q_l}$. Even if $\Pr(\delta_Q = 1|Y)$ is large, the data may suggest that certain coefficients are small since $\delta_Q = 1$ does not imply that the coefficients must be far from zero. This means that if some components are not important, the posteriors for the corresponding regression coefficients will be near zero. If the questions are only being asked informally, then it seems reasonable to use this approach, rather than modify any prior beliefs. See Section 7.3.2 for an example of this approach and further discussion.

Not only does this approach reduce the size of the total model space, but it makes headway in dealing with the pitfalls of multiple comparisons. When a group of variables is not active, but there are many of them, the chances are large that one variable will appear active because of random variation. By only considering the importance of the group, the chances of making such an error are reduced, because either a large single effect or several medium sized effects will be needed to conclude that the group is active.

This type of relation can of course be generalized and relaxed in many ways. The assumption that there is only a single term δ_Q associated with a group of terms could be relaxed to the case where each predictor has its own "importance variable" δ_i. All possible values of $(\delta_{Q_1}, \delta_{Q_2}, ..., \delta_{Q_l})$ will now have positive probability, but only the events $(\delta_{Q_1}, \delta_{Q_2}, ..., \delta_{Q_l}) \in \{(0, 0, ..., 0), (1, 1, ..., 1)\}$ will have probability greater than some small value ε.

5.4 Competing Predictors

Thus far, all the relations considered between predictors have been constructive. That is, when certain terms are active there is a chance that other terms will also be active. In this section a more competitive or exclusive relation between predictors is considered, namely an either/or relation. Certain variables may be mutually exclusive from other variables, either for practical or scientific reasons. For example, defects may be caused by one of two stages in a process, but it might be suspected that only one stage was the culprit. Without information to suggest which stage is to blame this is a problem of competing predictors. Another application is related to the choice of transformations for predictor variables. The response may be related linearly to the variable A or some transformed version of that variable such as $\sqrt{A}, \log A$ or A^{-1}. Unless the model is quite complex, only one of these predictors would be included in the model. If only a few transformations of A are considered, then this problem may be viewed as a variable selection problem with each different transformation as a competing predictor.

In both cases (transformations and mutually exclusive predictors), certain predictors must not be active in the same model. Consider, for example predictors A and B, and suppose it is not clear whether A or $\sqrt{A}$ should be used. Five models involving the three terms $A, B, \sqrt{A}$ may be considered: the model of no effect, $\{\}$, and the models $\{A\}, \{\sqrt{A}\}, \{A, B\}$, and $\{\sqrt{A}, B\}$. The variables A and $\sqrt{A}$ are not allowed to appear in the same model. This belief may be expressed in the priors for $\boldsymbol{\delta}$. There are many ways to do this, and it will be important to select a prior that uses as much structure as possible. Assigning zero mass to all "impossible" models (i.e. those containing both A and $\sqrt{A}$) and re-normalizing the probabilities for all other models is not a good approach, since the re-normalization can drastically alter the probability structure for models not receiving zero mass (see the next section for further comments on reweighting).

A better way to capture this mutually exclusive structure is to use a mixture of models. Define k sets of indices M_1, M_2, ..., M_k for $\boldsymbol{\delta}$. Each set M_i indexes a subvector $\boldsymbol{\delta}_{M_I}$ of $\boldsymbol{\delta}$, and corresponds to one of the competing models. For the above example,

$$M_1 = \{A, B\} \Rightarrow \boldsymbol{\delta}_{M_1} = (\delta_A, \delta_B)$$

$$M_2 = \{\sqrt{A}, B\} \Rightarrow \boldsymbol{\delta}_{M_2} = (\delta_{\sqrt{A}}, \delta_B).$$

The model with only B active may be expressed in terms of either subvector, with the first element of $\boldsymbol{\delta}_{M_i}$ zero. Note that the indices in any set M_i represent two distinct groups: terms which compete with other terms and terms which are common to all models. In the example, $A \in M_1$ is a competing term, and $B \in M_1$ is a common term. If there are several competing terms in a set M_i, none of them may compete with each other. The complement of M_i, denoted $\overline{M}_i$, represents all terms which compete with the elements in M_i. So

$$\boldsymbol{\delta}_{M_i} \neq \mathbf{0} \Rightarrow \boldsymbol{\delta}_{\overline{M}_i} = \mathbf{0}.$$

A prior representing competing predictors may be expressed mathematically as a mixture of priors for the individual models, with mixture probabilities $p_1, \ldots p_k$ such that $\sum p_i = 1$:

$$\Pr(\boldsymbol{\delta}) = \sum_{i=1}^{k} \Pr(\boldsymbol{\delta}_{M_i}) I(\boldsymbol{\delta}_{\overline{M}_i} = 0) p_i. \tag{5.11}$$

Since $\sum_{\boldsymbol{\delta}_{M_i}} \Pr(\boldsymbol{\delta}_{M_i}) = 1$ for any fixed i, the total probability assigned to the model space is still 1. The mixture parameters $p_1, \ldots, p_k$ are the probabilities that each model is true; if there is little prior knowledge, $p_i = 1/k$ might be chosen. Note also that the event $\boldsymbol{\delta} = \mathbf{0}$ is common to all terms in the above summation, rather than occurring separately or in a single term. This is not problematic, since each of the terms is weighted by p_i. This formulation is intuitive, since it corresponds to the idea that conditional on a certain model M_i, the remaining unspecified elements of $\boldsymbol{\delta}$ (i.e. $\boldsymbol{\delta}_{M_i}$) are free to take on any possible value. That is, conditional on a model, the remaining values of $\boldsymbol{\delta}$ specify a complete set of submodels which include all possible values of $\boldsymbol{\delta}$ not specified by the conditioning on model M_i. The adaptation of this structure to include principles introduced in previous sections is straightforward, since terms of the form $\Pr(\boldsymbol{\delta}_{M_i})$ are completely arbitrary, and could utilize hierarchical structure, effect grouping, or any other structures one might wish to consider. Alternately, a competing group could form a small component of a larger model, provided the group was assumed independent of the rest of the model.

To clarify this further, consider the following example. Suppose that predictors A, B, and C are observed, and all two-way interactions between these three factors are considered. Values of another variable a are available, but models with both A and a make no sense. The variables a and A are competing predictors, as are the two-way interactions $\{aB, aC\}$, $\{AB, AC\}$. In the notation developed above, there are two competing models,

$$\begin{aligned} M_1 : A, B, C, AB, AC, BC \qquad & \overline{M}_1 : a, aB, aC \\ M_2 : a, B, C, aB, aC, BC \qquad & \overline{M}_2 : A, AB, AC. \end{aligned} \tag{5.12}$$

The prior for $\boldsymbol{\delta}$ may be written as a mixture of two models, one with terms not involving a, and the other not involving A. Since there are only two competing models, mixing probabilities p_1 and $1 - p_1$ will be used for models M_1 and M_2 respectively.

$$\Pr(\boldsymbol{\delta}) = \Pr(\boldsymbol{\delta}_{M_1}) I(\boldsymbol{\delta}_{\overline{M}_1} = 0) p_1 + \Pr(\boldsymbol{\delta}_{M_2}) I(\boldsymbol{\delta}_{\overline{M}_2} = 0)(1 - p_1) \tag{5.13}$$

Assuming the principles outlined in the previous section hold, the terms in (5.13) may be simplified. For example, the conditional independence principle and the inheritance principle imply

$$\begin{aligned} \Pr(\boldsymbol{\delta}_{M_1}) &= \Pr(\delta_A, \delta_B, \delta_C, \delta_{AB}, \delta_{AC}, \delta_{BC}) \\ &= \Pr(\delta_A)\Pr(\delta_B)\Pr(\delta_C)\Pr(\delta_{AB}|\delta_A, \delta_B)\Pr(\delta_{AC}|\delta_A, \delta_C)\Pr(\delta_{BC}|\delta_B, \delta_C) \end{aligned} \tag{5.14}$$

Once the same simplification is made for $\Pr(M_2)$, the prior on $\boldsymbol{\delta}$ in (5.13) may be expressed as

$$\begin{aligned} \Pr(\boldsymbol{\delta}) =& \Pr(\delta_B)\Pr(\delta_C)\Pr(\delta_{BC}|\delta_B, \delta_C) \times \\ & \{\Pr(\delta_A)\Pr(\delta_{AB}|\delta_A, \delta_B)\Pr(\delta_{AC}|\delta_A, \delta_C) I(\{\delta_a, \delta_{aB}, \delta_{aC}\} = \mathbf{0}) p_1 + \\ & \Pr(\delta_a)\Pr(\delta_{aB}|\delta_a, \delta_B)\Pr(\delta_{aC}|\delta_a, \delta_C) I(\{\delta_A, \delta_{AB}, \delta_C\} = \mathbf{0})(1 - p_1)\} \end{aligned} \tag{5.15}$$

Note that in (5.15) terms which are common to both models (here B, C, BC) may be factored out of the individual components of the equation. In this case, the conditional independence principle and the effect heredity principle allowed the terms to be factored out. This is generally the case, since conditional independence breaks up the probability into a product of terms, and inheritance removes any competing terms from probabilities for terms common to all models. In this case, if inheritance was not assumed, then the term $\Pr(\delta_{BC}|\delta_A, \delta_B, \delta_C)$ would still depend on δ_A and could not be factored.

Although the principles of conditional independence and inheritance are sufficient conditions for the factoring, they are not necessary. All that is required to factor out terms is independence between the terms in a competing group and terms in the common group. In this case, both of the sets of terms $\{A, AB, AC\}$ and $\{a, aB, aC\}$ must be independent of $\{B, C, BC\}$. Clearly the principles of conditional independence and inheritance together imply this weaker condition. This obvious result is of limited use, since in practice it seems less likely that the independence of groups would be assumed without also assuming conditional independence and inheritance.

An industrial example in Section 6.4 illustrates a situation in which competing predictors arise from aliasing relations.

5.5 Global Constraints on Models

This section considers a supplementary method for writing out the prior on $\boldsymbol{\delta}$ which is based on more global constraints. This is in contrast with previous methods, which described relations between individual terms. In general, suppose that some probability model (i.e., a prior) for $\boldsymbol{\delta}$, denoted by $f(\boldsymbol{\delta})$ has already been specified. This prior could be weighted according to some global property of $\boldsymbol{\delta}$, using a weighting function $w(\boldsymbol{\delta})$. That is, construct a weighted prior

$$f_w(\boldsymbol{\delta}) \propto w(\boldsymbol{\delta}) f(\boldsymbol{\delta}), \tag{5.16}$$

where $w(\boldsymbol{\delta})$ reweights according to some global property of $\boldsymbol{\delta}$.

The most obvious global property that $w(\boldsymbol{\delta})$ might represent is the number of active terms in the model, and this was the original motivation for using such a construction. For example,

$$w(\boldsymbol{\delta}) = I(|\boldsymbol{\delta}| \leq 10)$$

where the number of active terms is given by

$$|\boldsymbol{\delta}| = \sum_{i=1}^{p} \delta_i$$

would restrict attention to those models with fewer than ten active terms. GM suggest the prior

$$w(\boldsymbol{\delta}) = w_{|\boldsymbol{\delta}|} \binom{p}{|\boldsymbol{\delta}|}^{-1}$$

where

$$w_{|\boldsymbol{\delta}|} = \Pr(|\boldsymbol{\delta}| \text{ terms active}).$$

This is another example of weighting, in this case with $w(\boldsymbol{\delta})$ as given, and $f(\boldsymbol{\delta}) = \text{constant}$.

No matter how it is defined, the weight $w(\boldsymbol{\delta})$ will be discretely valued since $\boldsymbol{\delta}$ has a finite number of elements. This means that $w(\boldsymbol{\delta})$ may be viewed as defining a partition of the model space, and then reweighting each partition. Of course, interest will focus on simple partitions such as the examples given above. Variations on the number of terms might be used for reweighting - for example, the number of main effects, interactions or terms of a certain order might be used to define partitions for reweighting.

It is interesting to note that in general the weighted partition will not preserve the probability structure described in the previous sections. Consider a simple case involving main effects A, B, and C. Assume independence of δ_A, δ_B, and δ_C, and reweight $\boldsymbol{\delta} = (1,1,1)$ to have zero prior mass, then $\Pr(\delta_A = 1|\delta_B = \delta_C = 1) = 0$ but $\Pr(\delta_A = 1) \neq 0$. Note that conditional on any given partition defined by w, the structures and relations defined previously still hold. Also, it is clear that any strong restrictions (those which assign probability zero to certain models, such as strong heredity) will still hold. What will change is the probabilities of certain events (due to the reweighting), and independence assumptions. The extent of this change will depend on the unweighted probability of the partitions and the weighting function.

One question that should be asked is whether this approach is really necessary. If the priors are constructed properly using the suggestions in the previous sections, and the probabilities chosen properly, does this also take care of the type of global restrictions described here? In many situations this may be the case. It is recommended that the unweighted prior probabilities of each proposed partition be determined (either analytically or through some sampling approach) *before* any partitions are reweighted. Even when this is done, it may be believed that certain partitions are in fact impossible, and such reweighting may be justified. Examples in which this reweighting is a convenient tool include cases where there are more predictors than observed data points, and will be discussed in more detail in Section 6.3

5.6 Computational Details

Some computational details of the Gibbs sampling algorithm for variable selection with related predictors are given in this section. They fall under two broad categories: convergence of the Gibbs sampler and the sampling from the full conditional distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$.

5.6.1 Convergence

The general approach taken in generating samples is to utilize long runs of the Gibbs sampler, rather than many short runs. As discussed in Section 2.2.1, "short runs" implies that the final values of n separate chains with different starting values are used. The term "long runs" refers to the use of a single run to generate many samples, taken at evenly spaced intervals. In situations where many (say 100 or more) runs are needed to achieve convergence, long runs are more efficient because fewer samples need to be discarded.

The issue of determining convergence in real situations is a difficult one, and there is considerable literature on the subject. Gelman and Rubin (1992) and Geyer (1992) provide differing views on the merits of using several long runs to determine convergence. Geyer places more emphasis on the use of a single long chain, although sample based convergence diagnostics are not available for such an approach. To obtain sample based convergence diagnostics, Gelman and Rubin advocate the choice of several starting values in different regions of the parameter space, and then running the independent chains until they appear similar. There are numerous ways to measure this "similarity", including inspection of properties of marginal distributions such as quantiles and moments, or the examination of higher dimensional distributional properties.

In a discussion of the two papers mentioned above, Gelfand (1992) advocates methods which assess similarity of the entire vector $\boldsymbol{\theta}$ rather than univariate components of it. While this would be difficult for many situations, some multivariate measures of similarity can easily be applied to $\boldsymbol{\delta}$ in this problem. For example, the 30 most probable models could be identified for each independent sequence at various points in the sampling procedure, and the chains continued until the same models are in this set, and they have comparable probabilities. Once a ballpark number of iterations for convergence has been identified, a single long run with longer length will be generated for inference. In the case study in Chapter 7, it was determined that by 10000 iterations, the variability between two independent chains was acceptable. The prior and posterior analyses were based on a single long chain of 50000 cycles.

The above discussion of sample-based convergence diagnostics is pertinent for the variable selection problem, because the priors introduced in Chapter 5 are quite likely to make convergence of the algorithm much slower. GM (p. 885) point out that when there are several models with high probability and the probability of a transition from one model to the other is small, convergence may be slow. This will quite often be the case, since many of the relationships in Chapter 5 downweight or eliminate a large subset of the model space. By doing so, the number of paths between any given models is greatly reduced. For example, suppose that strong heredity holds, and two possible models are $\{A, B, AB\}$ and $\{C, D\}$. To move from the former to the latter, the AB term *must* be the first one dropped, since both A and B must be in a model containing AB. If all terms were independent, other paths from the first to the second model would be allowed.

This problem is especially acute in the case of competing predictors, where the only common ground shared by two competing models are the models with none of the competing terms active. So in order to make a transition from model M_1 to model M_2, all terms in M_1 which compete with terms in M_2 (namely the terms $\overline{M}_1$) must be inactive. If there are many terms, this could take some time.

5.6.2 Gibbs Sampling with Variable Selection

The general idea of a hierarchical variable selection prior on regression coefficients is applicable to many problems. In this section, three cases are considered - two different priors for linear regression and the ordinal problem. The variables used in some or all of these cases are

$$\begin{array}{ll} \mathbf{Z} & \text{vector of continuous responses} \\ \mathbf{Y} & \text{discretized version of } \mathbf{Z} \\ \sigma & \text{variance of elements of } \mathbf{Z} \\ \boldsymbol{\beta} & \text{regression coefficients} \\ \boldsymbol{\delta} & \text{variable selection vector} \\ \gamma & \text{cutpoints on a latent scale.} \end{array} \tag{5.17}$$

The full conditional distributions for the Gibbs algorithm may be determined from the joint distribution of the data and the parameters. For the linear problem, GM use a "nonconjugate" prior. The joint distribution is given by:

$$f(\mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\delta}, \sigma) = f(\mathbf{Z}|\boldsymbol{\beta}, \sigma) f(\boldsymbol{\beta}|\boldsymbol{\delta}) f(\sigma|\boldsymbol{\delta}) f(\boldsymbol{\delta}). \tag{5.18}$$

In George and McCulloch (1994), a "conjugate" prior is used, yielding

$$f(\mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\delta}, \sigma) = f(\mathbf{Z}|\boldsymbol{\beta}, \sigma) f(\boldsymbol{\beta}|\boldsymbol{\delta}, \sigma) f(\sigma|\boldsymbol{\delta}) f(\boldsymbol{\delta}). \tag{5.19}$$

In ordinal regression with a latent variable, the joint distribution is

$$f(\mathbf{Z}, \mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\delta}, \gamma) = f(\mathbf{Z}|\boldsymbol{\beta}) f(\boldsymbol{\beta}|\boldsymbol{\delta}) f(\boldsymbol{\delta}) f(\gamma) \mathcal{I}(\mathbf{Y}; \mathbf{Z}, \gamma), \tag{5.20}$$

where the indicator function $\mathcal{I}$ is one when $\mathbf{Y}$ agrees with $(\mathbf{Z}, \gamma)$, and zero otherwise. The factoring of (5.19) assumes only that $f(\mathbf{Z}|\boldsymbol{\beta}, \sigma, \boldsymbol{\delta})$ does not depend on $\boldsymbol{\delta}$. In (5.18), the additional assumption is made that $\boldsymbol{\beta}$ and σ are independent given $\boldsymbol{\delta}$. The factoring of (5.20) is by an argument similar to the ordinal case without variable selection, with $f(\boldsymbol{\beta}|\boldsymbol{\delta}) f(\boldsymbol{\delta})$ replacing $f(\boldsymbol{\beta})$.

In (5.18) - (5.20) distributions of the same syntactic form are assumed identical. This means that all distributions in (5.18) and (5.19) are identical, except for $f(\boldsymbol{\beta}|\boldsymbol{\delta}, \sigma)$ and $f(\boldsymbol{\beta}|\boldsymbol{\delta})$. The exact forms of distributions in (5.18) - (5.20) used here are

$$\begin{aligned} f(\mathbf{Z}|\boldsymbol{\beta}, \sigma) &= N(\mathbf{X}'\boldsymbol{\beta}, \sigma^2 I) \\ f(\mathbf{Z}|\boldsymbol{\beta}) &= f(\mathbf{Z}|\boldsymbol{\beta}, \sigma = 1) \\ f(\boldsymbol{\beta}|\boldsymbol{\delta}, \sigma) &= N(0, \sigma^2 \mathbf{D}_{\boldsymbol{\delta}} \mathbf{R} \mathbf{D}_{\boldsymbol{\delta}}) \text{ where } \mathbf{D} = \operatorname{diag}(c_1^{\delta_1} \tau_1, \ldots, c_p^{\delta_p} \tau_p) \\ f(\boldsymbol{\beta}|\boldsymbol{\delta}) &= f(\boldsymbol{\beta}|\boldsymbol{\delta}, \sigma = 1) \\ f(\sigma^2|\boldsymbol{\delta}) &= IG(\nu_{\boldsymbol{\delta}}/2, \nu_{\boldsymbol{\delta}} \lambda_{\boldsymbol{\delta}}/2) \\ f(\boldsymbol{\delta}) &= \text{an arbitrary multinomial on } 2^p \text{ values} \\ f(\gamma) &= \text{an arbitrary distribution such that } \gamma_1 < \ldots < \gamma_{J-1} \\ \mathcal{I} &= 1 \text{ if } (\mathbf{Y}, \mathbf{Z}, \gamma) \text{ agree, 0 otherwise} \end{aligned} \tag{5.21}$$

Before similarities and differences between (5.18) - (5.20) are discussed, observe that (5.20) is a simple extension of the ordinal problem without variable selection. Comparison of (5.20) with the original model (without $\boldsymbol{\delta}$),

$$f(\mathbf{Z}, \boldsymbol{\beta}, \gamma, \mathbf{Y}) = f(\mathbf{Z}|\boldsymbol{\beta}) f(\boldsymbol{\beta}) f(\gamma) \mathcal{I}(\mathbf{Y}; \mathbf{Z}, \gamma), \tag{2.9}$$

reveals that the only difference is that $f(\boldsymbol{\beta})$ is replaced by $f(\boldsymbol{\beta}|\boldsymbol{\delta})f(\boldsymbol{\delta})$. This means the Gibbs steps for $\mathbf{Z}$ and γ remain unchanged from the original case, and only the steps for sampling $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ change. The exact form of these sampling steps will be discussed below.

The Gibbs sampling steps for $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are common to all three models (5.18) - (5.20), and share some common forms. In all three cases, the draw for $\boldsymbol{\beta}$ is a simple multivariate normal generation. The conditional distribution for the conjugate case (the most general of the three) when $f(\boldsymbol{\beta}|\boldsymbol{\delta}, \sigma) = N(0, \sigma^2 \mathbf{D}_{\boldsymbol{\delta}} \mathbf{R} \mathbf{D}_{\boldsymbol{\delta}})$ is

$$f(\boldsymbol{\beta}|\text{all other variables}) \propto f(\mathbf{Z}|\boldsymbol{\beta}, \sigma) f(\boldsymbol{\beta}|\boldsymbol{\delta}, \sigma) \propto N(\sigma^{-2} \mathbf{A}_{\boldsymbol{\delta}} \mathbf{X}'\mathbf{Z}, \mathbf{A}_{\boldsymbol{\delta}}), \tag{5.22}$$

where

$$\mathbf{A}_{\boldsymbol{\delta}} = \sigma^{-2} (\mathbf{X}'\mathbf{X} + \mathbf{D}_{\boldsymbol{\delta}}^{-1} \mathbf{R}^{-1} \mathbf{D}_{\boldsymbol{\delta}}^{-1})^{-1}. \tag{5.23}$$

In the non-conjugate case (5.18), $f(\boldsymbol{\beta}|\boldsymbol{\delta}, \sigma)$ becomes $f(\boldsymbol{\beta}|\boldsymbol{\delta}) = N(0, \mathbf{D}_{\boldsymbol{\delta}} \mathbf{R} \mathbf{D}_{\boldsymbol{\delta}})$ and (5.22) holds with

$$\mathbf{A}_{\boldsymbol{\delta}} = (\sigma^{-2} \mathbf{X}'\mathbf{X} + \mathbf{D}_{\boldsymbol{\delta}}^{-1} \mathbf{R}^{-1} \mathbf{D}_{\boldsymbol{\delta}}^{-1})^{-1}. \tag{5.24}$$

Setting $\sigma = 1$ yields the full conditional distribution for the ordinal case. If $\mathbf{X}$ is orthogonal and $\mathbf{R}$ diagonal, then the inversions in (5.23) and (5.24) become trivial, and computation faster. In other cases, some numerical shortcuts are available - see George and McCulloch (1994) for details.

The draws of $\boldsymbol{\delta}$ are different for the three models, but share common structure. The three full conditional distributions for $\boldsymbol{\delta}$ are

$$\begin{aligned} f(\boldsymbol{\delta}|\mathbf{Z}, \boldsymbol{\beta}, \sigma) &\propto f(\boldsymbol{\beta}|\boldsymbol{\delta}) f(\sigma|\boldsymbol{\delta}) f(\boldsymbol{\delta}) \\ f(\boldsymbol{\delta}|\mathbf{Z}, \boldsymbol{\beta}, \sigma) &\propto f(\boldsymbol{\beta}|\boldsymbol{\delta}, \sigma) f(\sigma|\boldsymbol{\delta}) f(\boldsymbol{\delta}) \\ f(\boldsymbol{\delta}|\mathbf{Z}, \boldsymbol{\beta}, \mathbf{Y}, \gamma) &\propto f(\boldsymbol{\beta}|\boldsymbol{\delta}) f(\boldsymbol{\delta}), \end{aligned} \tag{5.25}$$

for the nonconjugate, conjugate, and ordinal models. As mentioned in Section 4.2, each element δ_i is sampled conditional on the remaining elements $\boldsymbol{\delta}_{(-i)}$ and other parameters. If $\boldsymbol{\delta}$ was sampled in one step, it would be necessary to enumerate the entire space of the $\boldsymbol{\delta}$ components. A compromise approach that considers sampling groups of $\boldsymbol{\delta}$ is discussed in George and McCulloch (1994).

To draw δ_i from its full conditional distribution, it is necessary to calculate the probability that $\delta_i = 1$, conditional on all other variables. First note that each full conditional distribution in (5.25) has the form

$$f(\delta_i|\boldsymbol{\delta}_{(-i)}, \text{all other variables }) \propto h(\delta_i, \boldsymbol{\delta}_{(-i)}) f(\delta_i, \boldsymbol{\delta}_{(-i)}),$$

where the dependence of h on other variables is suppressed. So for example, under the nonconjugate model,

$$h(\delta_i, \boldsymbol{\delta}_{(-i)}) = f(\boldsymbol{\beta}|\delta_i, \boldsymbol{\delta}_{(-i)}) f(\sigma|\delta_i, \boldsymbol{\delta}_{(-i)}).$$

Since δ_i may only assume values 0 and 1,

$$
\begin{aligned}
\Pr(\delta_i = 1|\boldsymbol{\delta}_{(-i)}, \text{other variables}) &= \frac{f(1,\boldsymbol{\delta}_{(-i)})h(1,\boldsymbol{\delta}_{(-i)})}{f(1,\boldsymbol{\delta}_{(-i)})h(1,\boldsymbol{\delta}_{(-i)}) + f(0,\boldsymbol{\delta}_{(-i)})h(0,\boldsymbol{\delta}_{(-i)})} \\
&= \frac{f(1,\boldsymbol{\delta}_{(-i)})}{f(1,\boldsymbol{\delta}_{(-i)}) + f(0,\boldsymbol{\delta}_{(-i)})[h(0,\boldsymbol{\delta}_{(-i)})/h(1,\boldsymbol{\delta}_{(-i)})]}.
\end{aligned} \tag{5.26}
$$

The function h is written as a ratio since it is assumed to be positive over the parameter space. This is not true of f, since certain models may have mass zero. A draw of δ_i requires evaluation of the joint prior of $\boldsymbol{\delta}$ at $(\delta_i = 1, \boldsymbol{\delta}_{(-i)})$ and $(\delta_i = 0, \boldsymbol{\delta}_{(-i)})$, and calculation of the ratio $h(0,\boldsymbol{\delta}_{(-i)})/h(1,\boldsymbol{\delta}_{(-i)})$. Evaluation of the joint prior of $\boldsymbol{\delta}$ can be up to an arbitrary constant which cancels in the numerator and denominator of (5.26).

Details for the calculation of $h(0,\boldsymbol{\delta}_{(-i)})/h(1,\boldsymbol{\delta}_{(-i)})$ are given under the assumption that $f(\sigma|\boldsymbol{\delta}) = f(\sigma)$. This assumption is used for all examples in GM and George and McCulloch (1994), and includes the ordinal algorithm with $\sigma = 1$. To include the case of grouped predictors, consider a comparison of $\boldsymbol{\delta}_1 = (\delta_{11}, \ldots, \delta_{1p})$ and $\boldsymbol{\delta}_0 = (\delta_{01}, \ldots, \delta_{0p})$, where these vectors are identical except for a set S of subscripts. That is,

$$
\begin{aligned}
\delta_{1i} = \delta_{0i} \qquad & i \notin S \\
\delta_{1i} = 1; \delta_{0i} = 0 \qquad & i \in S.
\end{aligned} \tag{5.27}
$$

Even if there is grouping, assume that $\boldsymbol{\delta}$ is the same length as $\boldsymbol{\beta}$, with repeated elements. Also assume that each β_i has its own c_i and τ_i. The notation $\mathbf{D}_0$ and $\mathbf{D}_1$ is used to denote $\mathbf{D}$'s corresponding to $\boldsymbol{\delta}_0$ and $\boldsymbol{\delta}_1$. Then

$$
\mathbf{D}_k = \mathrm{diag}(c_1^{\delta_{k1}}\tau_1, \ldots, c_p^{\delta_{kp}}\tau_p) = \mathrm{diag}(d_{k1}^{-1}, \ldots, d_{kp}^{-1}) \qquad \text{for } k = 0, 1.
$$

The inverses make subsequent calculations cleaner. The same logic is used for $\mathbf{R}^{-1}$, the inverted prior covariance of $\boldsymbol{\beta}$, which is assumed to have ijth element r_{ij}. The following relations from (5.27) will be useful:

$$
\begin{aligned}
d_{1i} = d_{0i} \qquad & i \notin S \\
d_{1i} \neq d_{0i} \qquad & i \in S.
\end{aligned} \tag{5.28}
$$

The ratio of the normal densities can then be written

$$
\begin{aligned}
h(0,\boldsymbol{\delta}_{(-i)})/h(1,\boldsymbol{\delta}_{(-i)}) &= f(\boldsymbol{\beta}|\boldsymbol{\delta}_0)/f(\boldsymbol{\beta}|\boldsymbol{\delta}_1) \\
&= \frac{|\sigma^2\mathbf{D}_1\mathbf{R}\mathbf{D}_1|^{-1/2}}{|\sigma^2\mathbf{D}_0\mathbf{R}\mathbf{D}_0|^{-1/2}} \times \\
&\qquad \exp(-\frac{1}{2}\{\boldsymbol{\beta}'[\sigma^2\mathbf{D}_1\mathbf{R}\mathbf{D}_1]^{-1}\boldsymbol{\beta} - \boldsymbol{\beta}'[\sigma^2\mathbf{D}_0\mathbf{R}\mathbf{D}_0]^{-1}\boldsymbol{\beta}\}) \\
&= \frac{|\mathbf{D}_0|}{|\mathbf{D}_1|}\exp(-\frac{1}{2\sigma^2}\boldsymbol{\beta}'[\mathbf{D}_1^{-1}\mathbf{R}^{-1}\mathbf{D}_1^{-1} - \mathbf{D}_0^{-1}\mathbf{R}^{-1}\mathbf{D}_0^{-1}]\boldsymbol{\beta})
\end{aligned} \tag{5.29}
$$

The matrix expression in the argument of the exponent has many zero elements and other structures. Using the relations in (5.28), its ijth element is

$$\begin{aligned}
(\mathbf{M})_{ij} &= (\mathbf{D}_1^{-1}\mathbf{R}^{-1}\mathbf{D}_1^{-1} - \mathbf{D}_0^{-1}\mathbf{R}^{-1}\mathbf{D}_0^{-1})_{ij} \\
&= r_{ij}(d_{0i}d_{0j} - d_{1i}d_{1j}) \\
&= r_{ij} \times \begin{cases} 0 & \text{for } i \notin S, j \notin S \\ d_{.j}(d_{0i} - d_{1i}) & \text{for } i \in S, j \notin S \\ d_{.i}(d_{0j} - d_{1j}) & \text{for } i \notin S, j \in S \\ d_{0i}d_{0j} - d_{1i}d_{1j} & \text{for } i \in S, j \in S \end{cases}
\end{aligned} \tag{5.30}$$

The "." subscript emphasizes that the corresponding d is the same in both $\mathbf{D}_0$ and $\mathbf{D}_1$. Premultiplication of $\mathbf{M}$ by $\boldsymbol{\beta}$ in (5.29) yields a vector whose jth element is

$$(\boldsymbol{\beta}'\mathbf{M})_j = \begin{cases} \sum_{i \in S} \beta_i r_{ij} d_{.j}(d_{0i} - d_{1i}) & j \notin S \\ \sum_{i \notin S} \beta_i r_{ij} d_{.i}(d_{0j} - d_{1j}) + \sum_{i \in S} \beta_i r_{ij}(d_{0i}d_{0j} - d_{1i}d_{1j}) & j \in S \end{cases}. \tag{5.31}$$

Post-multiplication of the resultant vector by $\boldsymbol{\beta}$ gives

$$\begin{aligned}
\boldsymbol{\beta}'\mathbf{M}\boldsymbol{\beta} = \sum_{j \notin S} \beta_j \sum_{i \in S} \beta_i r_{ij} d_{.j}(d_{0i} - d_{1i}) &+ \sum_{j \in S} \beta_j \sum_{i \notin S} \beta_i r_{ij} d_{.i}(d_{0j} - d_{1j}) \\
&+ \sum_{j \in S} \beta_j \sum_{i \in S} \beta_i r_{ij}(d_{0i}d_{0j} - d_{1i}d_{1j}).
\end{aligned} \tag{5.32}$$

This may appear complicated, but there are far fewer calculations than multiplying entire matrices. In practice, the two types of sets considered will be $S = \{i\}$ for drawing δ_i and $S = \{i_1, , \ldots, i_k\}$ for drawing δ_i which corresponds to the k grouped terms $\beta_{i_1}, \ldots \beta_{i_k}$.

Further simplifications are possible. If $\mathbf{R}^{-1} = \mathrm{diag}(r_1, \ldots, r_p)$, then all r_{ij} terms above are zero for $i \neq j$, giving:

$$\begin{aligned}
\boldsymbol{\beta}'\mathbf{M}\boldsymbol{\beta} &= \sum_{i \in S} \beta_i^2 r_i (d_{0i}^2 - d_{1i}^2) \\
&= \sum_{i \in S} \beta_i^2 r_i (\tau_i^{-2} - (c_i\tau_i)^{-2}) \\
&= \sum_{i \in S} \beta_i^2 r_i (1 - c_i^{-2})/\tau_i^2
\end{aligned} \tag{5.33}$$

There are further simplifications if r_i, c_i, τ_i do not vary over subscript i, which can occur in some situations. The calculations above are intended to illustrate some of the shortcuts used in the current implementation, and also indicate the sort of simplifications that may be made in calculation of components of the full conditional distribution for $\boldsymbol{\delta}$.

The other component of the full conditional distribution of $\boldsymbol{\delta}$ to be calculated is the prior, namely $f(\delta_i = 0, \boldsymbol{\delta}_{(-i)})$ and $f(\delta_i = 1, \boldsymbol{\delta}_{(-i)})$. The form of $f(\boldsymbol{\delta})$ is the same for conjugate, nonconjugate, and ordinal models, so the following applies to all three cases.

The calculations necessary will depend on the types of relations used in the prior $f(\boldsymbol{\delta})$. Most of the forms in Chapter 5 are simple since there is often considerable factoring. The most attention will be given for inheritance relations, grouping and global constraints, since these are the ones currently implemented. Grouping adds no complications, since it does not influence relations between elements of $\boldsymbol{\delta}$. Inheritance and conditional independence (i.e. hierarchical polynomials and interactions) allow the prior for $\boldsymbol{\delta}$ to be factored into as many terms as there are elements of $\boldsymbol{\delta}$ (see equation (5.10)). Every distinct element $\delta_i \in \boldsymbol{\delta}$ has its own term of the form $\Pr(\delta_i|\mathcal{P}(\delta_i))$ where $\mathcal{P}(\delta_i)$ represents all parents of δ_i. In calculating $f(\delta_i = 0, \boldsymbol{\delta}_{(-i)})$ and $f(\delta_i = 1, \boldsymbol{\delta}_{(-i)})$ up to a normalizing constant for a specific element δ_i, it is only necessary to identify those factored terms of $f(\boldsymbol{\delta})$ which contain δ_i. The probability $\Pr(\delta_i|\mathcal{P}(\delta_i))$ and all other probabilities that condition on δ_i are thus factored out.

As the previous paragraph indicates, several data structures are needed to calculate these Bernoulli probabilities. For a specific δ_i the algorithm requires:

1. A list of the parents, $\mathcal{P}(\delta_i)$
2. A list of the children, $\{\delta_j|\delta_i \in \mathcal{P}(\delta_j)\}$
3. The conditional distribution $\Pr(\delta_k|\mathcal{P}(\delta_k))$ of the children in 2, and for $k = i$.

With these data structures, an algorithm for calculation of $f(\delta_i, \boldsymbol{\delta}_{(-i)})$ is straightforward to construct. As a simple example, consider a model with terms A, B, AB, A^2, and A^2B, and assume conditional independence, immediate inheritance, and strong heredity. The density of $\boldsymbol{\delta}$ may be written

$$\Pr(\boldsymbol{\delta}) = \Pr(\delta_A)\Pr(\delta_B)\Pr(\delta_{A^2}|\delta_A)\Pr(\delta_{AB}|\delta_A, \delta_B)\Pr(\delta_{A^2B}|\delta_{A^2}, \delta_{AB}).$$

The term A has no parents, and children AB and A^2. Thus

$$\Pr(\delta_A = 1|\delta_B, \delta_{AB}, \delta_{A^2}, \delta_{A^2B}) \propto \Pr(\delta_A = 1)\Pr(\delta_{A^2}|\delta_A = 1)\Pr(\delta_{AB}|\delta_A = 1, \delta_B).$$

The term AB has parents A and B, and child A^2B, so

$$\Pr(\delta_{AB} = 1|\delta_A, \delta_B, \delta_{A^2}, \delta_{A^2B}) \propto \Pr(\delta_{AB} = 1|\delta_A, \delta_B)\Pr(\delta_{A^2B}|\delta_{A^2}, \delta_{AB} = 1).$$

In both these cases, parents and children of a term were identified, and conditional probabilities for the term and its children evaluated.

The calculations for global constraints can be done after performing the above calculations, and the results multiplied, since global constraints are of the factored form $w(\boldsymbol{\delta})f(\boldsymbol{\delta})$. The currently implemented constraints reweight models depending on the number of active terms. For example, models with more than 10 active first order terms active could be made impossible by reweighting them to have no mass. This calculation of Bernoulli probabilities for δ_i amounts to counting the number of active terms for the situations $\delta_i = 0$ and 1, and reweighting based on those two counts.

Competing predictors (see equation (5.11)) would not require much further modification. It will be necessary to identify which of the competing models is active, or if the intersection of them is active (i.e. none of the terms which compete are active). If a single model is active, the algorithm is as before, simply applied to that model. If the intersection is active, all the densities of the competing models must be added together, which will mean much more work. The reason for this additional effort is that the usual proportionality constants must be calculated for each element in the mixture (5.11).

Miscellaneous Details

Starting values of $\boldsymbol{\delta} = (1, \ldots, 1)$ are used, although the values $(0, \ldots, 0)$ might be used for a second run to asses convergence. In practice, $|\boldsymbol{\delta}|$ falls to a small size quite quickly, implying that the starting value is of little importance. The sampling order of the δ_i's is randomized for each loop to facilitate better mixing (as suggested in GM).

The values of c_i and τ_i are chosen with the suggestions of George, McCulloch and Tsay (1993) in mind. They point out that with a transformed response (in this case via a link function), the effects of terms are linear on the underlying continuous scale, and not on the transformed scale. This means that the effect of an additive shift Δ added to some location μ on the latent scale will produce a shift in cumulative probability of $\Phi(\mu + \Delta)$. The magnitude of this shift depends on μ, so the best that can be achieved is to limit the largest possible effect on the probability scale. This problem applies to the choice of c_i and τ_i because they are typically chosen so that the magnitude of a negligible effect is $\sqrt{\tau_i}$ and the magnitude of an important effect is $\sqrt{\tau_i c_i}$. Obviously, the meaning of a negligible effect is dependent on the location, and statements can only be made about the largest size of a negligible effect. Typical values (for an orthogonal design $\mathbf{X}$ scaled such that $\mathbf{X}'\mathbf{X}/n = I$) would be $\tau = .1, c = 10$, yielding standard deviations for β_i of 0.1 and 1.

Chapter 6

Applications to Industrial Experiments

6.1 Robust Design

As discussed in Section 2.3.3, an exciting industrial application of designed experiments is Taguchi's "Robust Design" experiments. In such studies, predictors are divided into two general classes - those which are easy to control during regular production (control factors), and those which are not (noise factors). The goal of such studies is to choose control factor settings which desensitize the response to uncontrollable variation in the noise factors. In order to desensitize the response to the noise factors, it will be necessary to identify interactions between control and noise factors (see Shoemaker, Tsui and Wu, 1991). Quite often, these effects will be estimated, even if they are not large. This special interest in certain types of interactions may be incorporated into the variable selection prior using the techniques outlined in Chapter 5.

The exact way in which this is done is a matter of choice - several general recommendations are presented here. The letters C, N and CN are used to represent generic control factors, noise factors, and control by noise interactions.

- Use the weak heredity principle for $C \times N$ interactions instead of strong heredity. This allows $C \times N$ interactions more opportunity to be in the model, while retaining some of the structure and interpretability of hierarchically ordered models. A concrete example is given in Shoemaker, Tsui and Wu (1991), where the control by control by noise interaction AHM_3 is included even though only main effect H is large.

- Set $\Pr(\delta_{CN} = 1 | \delta_C = 0, \delta_N = 0) = \varepsilon > 0$ to allow interactions to enter into the model without main effects if the data provides substantial evidence.

- Use larger prior probabilities for the event $\delta_{CN} = 1$ than for other interactions. These probabilities may even be made as large or larger than those for main effects. This reflects the belief that there is more interest in estimation of control by noise interactions than control by control or noise by noise interactions. In fact, some practitioners would maintain that control by noise interactions are considered to be of the same order as control main effects. That is, in calculation of the order of a term, noise factors should not be counted. By increasing the probabilities that such terms are active while maintaining some hierarchical structure, a preference towards control by noise interactions is indicated. At the same time the hierarchical structure is maintained, meaning that interactions with active parents are more likely.

- Experiment with different values of τ_{CN} and c_{CN}, the parameters in (4.2). These define what is meant by a "large" effect, and in the case of $C \times N$ interactions an effect may be considered important even if it has a smaller magnitude than some other effects.

6.2 Non-Interacting Subsystems

Many industrial processes are quite complex, and involve different stages. For example, in a gasoline mixing process there are process variables that determine the qualities of the chemicals used, and there are also mixing variables that determine the recipe for producing the final product. In a well managed process of this type, the two different subsystems are unlikely to interact with one another. That is, an interaction between a process variable (P, say) and a mixture variable (M say) is unlikely to influence the response.

This belief may be represented in the prior for $\boldsymbol{\delta}$ by choosing $\Pr(\delta_{MP} = 1|\delta_M, \delta_P)$ to be $(0, 0, 0, \varepsilon)$ for the four values of $(\delta_M, \delta_P) \in \{(0,0), (0,1), (1,0), (1,1)\}$. The probability ε would be much smaller (perhaps by an order of magnitude) than similar probabilities for interactions between two process variables. By adding this constraint to the model selection process, it may be possible to reduce the number of candidate models substantially (or at least concentrate prior mass on fewer models). The gains will be largest when there are many possible interactions between the two subsystems which are a priori assumed unlikely to be important. In the case of a designed experiment, if these interactions are unlikely the design may not allocate any degrees of freedom for their estimation, and the model selection prior will not be needed. This type of prior will be most useful when interactions are correlated, either due to complex aliasing (see for example Hamada and Wu, 1992) or simply an observational study. Here, many interactions could be considered because of the correlations, so the prior assumption that certain interactions are not likely to be active reduces model complexity. This is more flexible than assuming an effect is small, since there is still a small probability that the interaction could be active.

This application extends far beyond industrial cases to include any problem in which many interactions could be entertained, and prior knowledge suggests that only a small subset have a large chance of being important. For example, in medical studies, a drug with a certain chemical makeup might be known to cause no side effects except when taken with alcohol. Prior probabilities can reflect the fact that the chance of an important alcohol by drug interaction is much larger than for any other interaction involving alcohol. A strong heredity structure is important here because it restricts the search to models that can be explained to practitioners. The observational nature of many medical studies makes them prime candidates for this approach because of the correlations between predictors and interactions.

6.3 Oversaturated Designs

Most of the examples explored in this thesis are saturated or nearly saturated. That is, the models fit to the data use all or almost all of the degrees of freedom available. The (often high) cost of carrying out each run is the main cause for such extensive use of the available data. In some situations, such as screening experiments where run costs are especially high or a large number of variables and/or interactions are under consideration, oversaturated designs may be used. As the name implies (the term "supersaturated" is also used), these designs have more predictor variables than degrees of freedom (say $n-1$). Recently there has been considerable interest in the construction (see, for example Wu, 1993) and analysis of data from such designs, indicated by Lin (1993) and others. Central to the analysis of data from such experiments are methods for identification of predictor subsets with fewer than n elements. Typically automated model selection techniques such as stepwise regression (or algorithms specially adapted to the designs - see Hamada and Wu (1992) for an example) have been employed.

In the current variable selection framework, the idea is to allocate only a small amount of prior mass to large models. There are two ways to accomplish this. The less drastic approach is to apply the sparsity principle and assume that each term is active with relatively small probability (either marginally or conditionally on parents, if the term has any). Box and Meyer's (1986) original application was in the context of saturated fractional factorial designs, and the motivation here is similar. They conduct a survey of experiments and suggest the use of $\Pr(\delta_i = 1)$ in the range $(0.1, 0.5)$ is reasonable. In the case of oversaturated designs, one might wish to favor the lower part of this range.

The second and more drastic way to put small prior mass on large models is to use the reweighting priors of Section 5.5. Weighting large models to have zero mass restricts the variable selection procedure to search only over models with $n-1$ terms or less. This idea of considering only models with less terms than observations was presented by Booth and Cox (1962), and more explicitly in Wu (1993).

As mentioned in the section on global constraints, it is recommended that before a subset (in this case all models with more than $n-1$ active terms) of the model space is downweighted, the probability of that subset be calculated. By doing this, priors that have abrupt changes in mass over different subsets will be avoided.

6.4 Experimental Mistakes and Almost Complete Aliasing

When designed experiments are run, accidents which destroy the orthogonal nature of the design occur frequently. Two common sources of error are missed runs and improperly set design levels. An example of the former might be a part failing during an experimental run, making it impossible to measure any response for that part. The latter problem is also quite common; levels of the design factors can easily be set at improper levels. For example in a two-level design, the levels might be reversed. Machine drift, misread design sheets, and any other number of possible errors could also lead to the factors being set at different levels.

In this section, the focus will be on fractional factorial designs in which there are small aberrations of the type described above. Typically a single run or a small number of runs will be affected, and either be unusable or involve nonstandard settings. In either case, the resulting design will be non-orthogonal. In the first case (missed runs), the problem is simple since the aliasing patterns are still the same. The correlations between the terms will be small but nonzero, meaning that numeric techniques that rely on orthogonal arrays will need modification. However, in comparison to the general regression problem, missed runs present no large challenges.

The second difficulty (improper levels) is more challenging, because the aliasing structures in the orthogonal array become complex. The fractional nature of the original design implies that many main effects and/or interactions will be involved in aliasing patterns. Or at least, they would be, if the experiment had been carried out as planned. Some of the terms in the model which would have had correlations of ± 1 now have correlations that are close to but not equal ± 1. So in theory, these terms could all be included in the model, and variable selection used to identify the most promising subsets.

It is evident that when the terms fall into groups that have high within-group correlations, it would be sensible to consider only one member of the group in the model at a time. For example one might use engineering considerations to identify a single factor in a group that is most likely to be important, or most likely to have a meaningful interpretation in the model. Then variable selection could be carried out with the restriction that only the identified variable in that group could enter the model - the other terms would be prohibited because of their high correlations. In this way the model selection process can be modified to take into account the information available from the planned design, which is assumed to be not that different from the design that actually was run.

If practical considerations are not sufficient to choose a single term from a highly correlated group, then the competing terms priors discussed in Section 5.4 might be used. In this application, each term in an almost aliased group would compete with all other distinct terms in the group - that is, only one term in a group would be allowed to be active. The distinction with the approach which uses subject matter knowledge is that no information is assumed to be known about which of the terms is active. The data may not aid this choice much, but if there is relevant information, then this method will pick it up.

One important advantage of this approach is that it automatically accounts for the correlation structure of the imperfect design, and will carry out automatic variable selection which acknowledges this structure.

To clarify the application of this technique and also illustrate the incorporation of interaction-main effect relations, consider the following example. Suppose a 16-run 2^{9-5}_{III} design is planned, with the nine main effects labeled $1, 2, \ldots, 9$, and the generating relations (taken from Table 12.15 of Box, Hunter, and Hunter (1978)) $I = 1235 = 2346 = 1347 = 1248 = 12349$. If all third and higher order interactions are assumed to be negligible, the design has the aliasing pattern:

$$\begin{array}{ll} 1 = 69 & 12 = 35 = 48 = 67 \\ 2 = 79 & 13 = 25 = 47 = 68 \\ 3 = 89 & 14 = 28 = 37 = 56 \\ 4 = 59 & 15 = 23 = 46 = 78 \\ 5 = 49 & 17 = 26 = 34 = 58 \\ 6 = 19 & 18 = 24 = 36 = 57 \\ 7 = 29 & \\ 8 = 39 & \\ 9 = 16 = 27 = 38 = 45 & \end{array} \tag{6.1}$$

It will be assumed that the interactions $16 = 27 = 38 = 45$ are negligible, since they are aliased with the main effect 9. If the design is run as planned, then aliased terms have correlations of 1. Suppose instead that the sign of variable 1 in the first run is changed. The design is no longer orthogonal, and there are partial correlations. Let x and y be two different terms, and use the notation $1 \in x$ to indicate that the term x involves the variable 1. Terms in the experiment as it was run will be denoted $\tilde{x}$ and $\tilde{y}$. The correlations in the new design are given by:

$$\mathrm{cor}(\tilde{x}, \tilde{y}) = \begin{cases} 0.88 & \text{if } 1 \in x, 1 \notin y, \mathrm{cor}(x, y) = 1 \\ \pm 0.13 & \text{if } 1 \in x, 1 \notin y, \mathrm{cor}(x, y) = 0 \\ \pm 0.02 & \text{if } 1 \in x, 1 \in y \\ \mathrm{cor}(x, y) & \text{otherwise} \end{cases} \tag{6.2}$$

Of course, the only correlations which change are those involving the variable with the reversed sign. Note that because of the defining relations of this specific design, any two terms both involving 1 are uncorrelated, and thus there is no correlation condition in the third line of (6.2).

Competing groups will involve those terms with correlations of 0.88, since the other relations are small enough to allow both terms to simultaneously enter the model. The correlations of 0.88 influence the aliasing patterns in (6.1), giving the modified relations

$$\begin{array}{ll} \mathrm{cor}(1,69) = 0.88 & \\ \mathrm{cor}(16,9) = 0.88, & 9 = 27 = 45 = 38 \\ \mathrm{cor}(12,35) = 0.88, & 35 = 48 = 67 \\ \mathrm{cor}(13,25) = 0.88, & 25 = 47 = 68 \\ \mathrm{cor}(14,28) = 0.88, & 28 = 37 = 56 \\ \mathrm{cor}(15,23) = 0.88, & 23 = 46 = 78 \\ \mathrm{cor}(17,34) = 0.88, & 26 = 34 = 58 \\ \mathrm{cor}(18,24) = 0.88, & 24 = 36 = 57 \end{array} \tag{6.3}$$

There are quite a few competing terms in (6.3). Each of the aliased groups is now subdivided into two competing terms - the variable containing a 1, and the remaining terms, which are indistinguishable. Assume that the interactions 69 and 16 are negligible, since they are aliased with main effects (although they could be used - see comments after the example). The six groups of competing terms (12 vs. 35, 13 vs. 25, 14 vs. 28, 15 vs. 23, 17 vs. 34, and 18 vs. 24) can all act independently, yielding a total of $2^6 = 64$ competing models. All of these models contain the common main effect terms 1,2,...,9, and one of each of the pairs of competing interaction terms. Assuming conditional independence and heredity, the prior on $\boldsymbol{\delta}$ is given by

$$\begin{aligned} \Pr(\boldsymbol{\delta}) &= \sum_{i=1}^{64} \Pr(\boldsymbol{\delta}_{M_i}) I(\boldsymbol{\delta}_{\overline{M}_i} = 0) p_i \qquad (6.4) \\ &= \Pr(\delta_1)\dots\Pr(\delta_9) \sum_{i=1}^{64} \Pr(\delta_{A(i)}|\boldsymbol{\delta}_{\mathcal{P}(A(i))})\dots\Pr(\delta_{F(i)}|\boldsymbol{\delta}_{\mathcal{P}(F(i))}) I(\boldsymbol{\delta}_{\overline{M}_i} = 0) p_i, \end{aligned}$$

where $A(i)$ is equal to either 12 or 35, and $\mathcal{P}(A(i))$ are the parents of $A(i)$, so if $A(i) = 12$ then $\mathcal{P}(A(i)) = \{1,2\}$. The same structure is used for $B,\dots,F$. The mixing probabilities may be chosen arbitrarily provided $\sum_{i=1}^{64} p_i = 1$. If models are assumed independent, the mixing probabilities are given by

$$p_i = \Pr(M_i) = p_{12}^{I_{12}(i)}(1-p_{12})^{(1-I_{12}(i))} p_{13}^{I_{13}(i)}(1-p_{13})^{(1-I_{13}(i))}\dots p_{18}^{I_{18}(i)}(1-p_{18})^{(1-I_{18}(i))}, \tag{6.5}$$

where p_{12} is the marginal probability that the model involves the variable 12, $1 - p_{12} = p_{35}$, and I_{12} is an indicator function for whether the variable 12 is in model M_i. Note that the mass associated with models for which $\delta_{12} = \delta_{35} = 0$ is split between p_{12} and $1 - p_{12}$. Typically, equal mass would be assigned to all of the competing models in the absence of additional information. In the above formulation, this corresponds to $\delta_{12} = 0.5$, and the same for the other five marginal probabilities.

The structure of this prior may seem quite complicated, but in fact it is not unmanageable. In the implementation developed here, the Gibbs sampling algorithm samples each element of $\boldsymbol{\delta}$ conditional on all the others, so all but one or two of the terms in (6.5) are zero, and sampling from full conditional priors is not more difficult than for other relations developed. It is also interesting to note that this type of example results in one of the more complicated structures for competing predictors, since if the predictors were competing because of reasons other than the design, then there would likely be fewer in competition.

In this example, it was assumed that the interactions 16 and 19 were negligible because they were aliased with main effects. An alternative would be to assign smaller prior probability to models which favor the interaction over the competing main effect. It is still conceivable that the interaction could be favored - for example, if main effects 1 and 6 appeared important, and the data favored the competing predictor 16 over 9, then a model with 16 instead of 9 might be entertained.

In this general type of problem, some care should be taken in using competing term models, since only a small number of observations are used to differentiate competing terms. In the example above, only one of the 16 runs has a mistake in it, so evidence in favor of any of the competing terms will come solely from this run. As mentioned earlier in this section, subject area knowledge should be used in identifying favored terms, and only in the absence of such knowledge should the competing predictors approach be used.

6.5 Optimization After Model Identification

This section does not outline an application; rather it is a note on the application of the optimization procedures of Section 2.3.3 in conjunction with variable selection. The former approach will be adopted with one small modification. Once a model is identified, the methods in Section 2.2.2 will be used to re-fit that specific model, with no variable selection parameters. Optimization would proceed as in Section 2.3.3.

Note that strictly speaking, if the priors represent all knowledge before the experiment, then the real posterior is a mixture of all models with mixing probabilities given by the posterior for $\boldsymbol{\delta}$. However, the original purpose of model selection was to identify a single model as an approximate description of the process. If all models were averaged together, it would be no improvement over just optimizing over the full model. Consequently, the choice is made to bias the model in the hopes that it will reduce the variance of predictions. The former averaging approach is considered by Madigan and Raftery (1991).

Chapter 7

Variable Selection Case Study

In this chapter, data from an industrial experiment will be analyzed using the variable selection techniques discussed in Section 4.2 and Chapter 5. After a brief description of the data (Section 7.1), attention will focus on the priors and posteriors for variable selection, namely distributions of $\boldsymbol{\delta}$ (Sections 7.2 and 7.3). Analysis will be based on prior and posterior samples, each of which contain 5000 draws obtained by taking every 10th observation from a single 50,000 cycle run of the Gibbs algorithm. For the sake of comparison, priors with different structures will be used, and comparisons made between them. Because even the structure in the priors can be complex, exploratory data analysis (EDA) will be carried out on *both* priors and posteriors. This EDA includes the introduction of several new plots designed to reveal aspects of the high dimensional prior and posterior samples of $\boldsymbol{\delta}$.

7.1 The Data

The data is from a study of the durability of solvent bonded connectors for use in automobile emission control systems. The data are presented and analyzed in Quinlan (1984). The goal of the experiment was to identify factors (and interactions of various orders) which influence the quality and durability of the solvent bond. A robust design experiment was used, with five control factors and three noise factors described in Table 7.1. The experimental design consisted of a product array of a nine run orthogonal noise array, and an eight run orthogonal control array, yielding 72 observations. Several response variables were observed; here the binary response indicating the survival (one) or failure (zero) of the part during the experiment will be modeled. The design and response data are given in Table 7.2.

The types of factors in the experiment and the levels that they have deserve some attention, as such structures will influence the analysis. The control factors A through D are at two levels each, and present little challenge. Control factor E is a continuous response set at four equally spaced levels, and will be coded as linear, quadratic and cubic contrasts (see Draper and Smith, 1981). These contrasts map the levels $(1, 2, 3, 4)$ to $(-3, -1, 1, 3)$, $(1, -1, -1, 1)$, and $(-1, 3, -3, 1)$, respectively, and yield an orthogonal estimation matrix.

Table 7.1 Factors and their levels, connector data

Control Factors

Name	Description	Levels
A	Tube material	(Nylon 6, Nylon 6/6)
B	Color Concentrate Type	(Olefinic/Polyamide)
C	Connector Type	(straight/elbow)
D	Applicator Design	(old/new)
E	Drying Time	(1s/2s/3s/4s)

Noise Factors

Name	Description	Levels
F	Tube Preheating	(Ambient/15min@350F/10min@400F)
G	Connector Turns	(0/1/2)
H	20 lb. Proof Loads	(0/5/10)

Table 7.2 Data from the connector experiment

					3	3	3	2	2	2	1	1	1	F
					3	2	1	3	2	1	3	2	1	G
A	B	C	D	E	1	3	2	2	1	3	3	2	1	H
1	1	1	1	1	0	1	1	0	1	1	1	1	1	
1	2	1	2	2	0	0	1	1	0	1	1	1	1	
1	2	2	1	3	1	1	1	0	1	1	1	1	1	
1	1	2	2	4	0	1	1	0	0	1	0	0	1	
2	2	2	2	1	0	0	0	0	0	0	0	0	1	
2	1	2	1	2	0	0	1	0	1	1	1	1	1	
2	1	1	2	3	0	0	1	0	0	0	0	1	1	
2	2	1	1	4	0	0	1	0	0	1	1	1	1	

Since there is a real continuous scale for the factor, these orthogonal contrasts will have meaning, and higher order terms will be associated with greater complexity. Similarly, the variables G and H are continuous with 3 levels, so linear $(-1, 0, 1)$ and quadratic $(1, -2, 1)$ contrasts are used. The factor F cannot be thought of as a simple univariate continuous predictor, since the levels consist of different temperature/time combinations, as well as an ambient heating method. If linear and quadratic contrasts are used, they have the interpretation of comparing $F = 1$ to $F = 3$, and $F = 2$ to the average of $F = 1$ and $F = 3$, respectively. These contrasts will be used because of their orthogonal properties, but methods for constructing more meaningful contrasts for interpretation will be discussed in Section 7.3.

With this coding, the control array allows for estimation of effects A, B, C, D, E, E^2, E^3, where E^2, E^3 represent quadratic and cubic contrasts. The noise array allows estimation of F_1, F_2, G, G^2, H, H^2, where F_1 and F_2 represent the two contrasts for factor F. Note that an additional two degrees of freedom for estimation are available from the noise array.

Since these degrees of freedom are partially aliased with many interactions involving the six terms to be estimated, they will not be estimated. The crossed array structure of the entire design means that in addition to the 13 terms mentioned above, all interactions between the control terms A, B, C, D, E, E^2, E^3 and the noise terms F_1, F_2, G, G^2, H, H^2 are estimable. These terms will be denoted by symbols such as AF_1, AF_2, AG, AG^2, AH, AH^2, ..., E^3F_1, E^3F_2, E^3G, E^3G^2, E^3H, E^3H^2. A proof of this estimability property is given in Shoemaker, Tsui, and Wu (1991), in proposition 1. In total, there are 55 terms in the regression model, ranging in order from main effects such as A to fifth order polynomials like E^3H^2. The polynomial relationships developed in Section 5.2 will be of use here to add structure to the model selection problem. The application of this and other prior structures for variable selection are considered in the next section.

7.2 Specification and Analysis of Variable Selection Priors

In this section, various priors for the model selection variables $\boldsymbol{\delta}$ will be discussed and examined. The examination will be somewhat detailed, because the priors contain considerable structure and can exert strong influence on the conclusions. For this thesis, a detailed analysis of priors is interesting for the sake of comparison. A more practical reason for analyzing priors is the need to perform a "reality check". Prior knowledge is often vague, and in this case very high dimensional, so looking at the same prior many different ways ensures that it actually is representative of prior knowledge.

Two specific types of "prior analysis" are of interest: a "structural" analysis of the relationships between the predictors, and analysis of the probability content of the priors. Before these analyses are given, an outline of the four priors to be compared is presented, along with some (reasonably general) justification for their choice.

7.2.1 Specification of Priors

All of the prior samples discussed below are generated by running the sampling algorithm in one long run of 50000 cycles, and storing every tenth sampled value of $\boldsymbol{\delta}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$. Details of the sampling based approach are given later in this section, under the heading "Exploring Probabilistic Aspects of Priors". The four priors are:

$\boxed{\text{IID}}$: Independent, identically distributed, $\Pr(\delta_i = 1) = 0.1$ (7.1)

$\boxed{\text{Independent}}$: Independent, terms of equal order identically distributed, $\Pr(\delta_i = 1) = \{0.25, 0.10, 0.05, 0.02, 0.01\}$ for orders 1, 2, 3, 4, 5. (7.2)

$\boxed{\text{Weak Heredity}}$: Dependent, with grouping (on terms involving F_1 and F_2), and (immediate) inheritance, $\Pr(\delta_A = 1) = 0.25$, $\Pr(\delta_{H^2} = 1|\delta_H) = (0, 0.5)$, $\Pr(\delta_{AG} = 1|\delta_A, \delta_G) = (0, 0.2, 0.2, 0.5)$. (7.3)

$\boxed{\text{Strong Heredity}}$: Identical to (7.3), except $\Pr(\delta_{AG} = 1|\delta_A, \delta_G) = (0, 0, 0, 0.5)$. (7.4)

Recall that the notation $\Pr(\delta_{H^2} = 1|\delta_H) = (p_1, p_2)$ and $\Pr(\delta_{AG} = 1|\delta_A, \delta_G) = (p_1, p_2, p_3, p_4)$ refers to

$$\Pr(\delta_{H^2} = 1|\delta_H) = \begin{cases} p_1 & \text{if } \delta_H = 0 \\ p_2 & \text{if } \delta_H = 1 \end{cases} \quad \text{and} \quad \Pr(\delta_{AG} = 1|\delta_A, \delta_G) = \begin{cases} p_1 & \text{if } (\delta_A, \delta_G) = (0,0) \\ p_2 & \text{if } (\delta_A, \delta_G) = (0,1) \\ p_3 & \text{if } (\delta_A, \delta_G) = (1,0) \\ p_4 & \text{if } (\delta_A, \delta_G) = (1,1) \end{cases}.$$

Independence priors are included for the sake of comparison, since independence is not an accurate representation of relationships between the predictors. The choices are intended to be comparable; the marginals for independent and weak heredity priors (7.2) and (7.3) are roughly equal, and the two independent priors have the same expected number of active terms. The heredity priors (7.3) and (7.4) are identical except for strong heredity on terms with two parents.

The grouping on terms involving F_1 and F_2 is as follows: a single element $\delta_{XF} \in \boldsymbol{\delta}$ represents a pair of terms XF_1, XF_2, where X is any other variable, including the intercept 1 (yielding terms F_1 and F_2). Further discussion of grouping and hierarchical structures is given in Section 7.2.2.

The sparsity principle and the fact that this is a screening experiment justifies the small probability of most terms being active. The large values for probabilities such as $\Pr(\delta_{AG} = 1|\delta_A = \delta_G = 1) = 0.5$ in (7.3) and (7.4) may be rationalized as follows. Since this probability is conditional on both main effects being active, and this happens with probability $0.25^2 = .0625$, so $\Pr(\delta_{AG} = 1)$ is actually quite small in the absence of information about δ_A or δ_G (in fact this is just the marginal probability, $0.5 \times 0.25^2 + 2 \times 0.2 \times 0.25 \times 0.75 = .1065$). In the case of strong heredity, such a marginal probability drops to 0.03125, quite a small value, despite the apparently large size of the 0.5 value in the conditional probability.

One common feature of all these priors is that they concentrate probability on a subspace of the entire model space. Independence priors accomplish this by stating that terms are unlikely to be active. The structured priors use small probabilities as well as structural relationships to further reduce (or entirely remove) prior mass from certain models. The heredity priors are perhaps unnecessarily harsh, and (as suggested in Section 5.1) probabilities of 0 could be relaxed to $\varepsilon > 0$. In what follows, the strict priors are considered because the patterns noticed will be more striking.

7.2.2 Exploring Structural Aspects of Priors

The term "structural relation" refers to dependence relations of the priors, and includes all of Chapter 5. Only dependent priors (7.3) and (7.4) are considered here, since independence needs no structural exploration. This section explains the grouping and heredity structural relations for the connector data, and illustrates their impact on the variable selection problem.

The grouping principle is applied to terms involving variables F_1 and F_2. By associating a single δ_{XF} with any pair of terms XF_1 and XF_2, only models with both or neither factor active are considered. That is, the question "Does temperature (or a temperature by X interaction) seem to make a difference ?" is asked, rather than immediately asking whether there are differences between pairs of settings (but see Section 7.3.1 for a way to decompose the effect of grouped terms once they are identified as a significant group). The implication for variable selection is that the number of terms is reduced from 55 to 47, or equivalently the model space is reduced to 0.39% of its original size (2^{55} to 2^{47}, a factor of $2^{-8} = 0.0039$). Heredity may further reduce the set of models, but those reductions will be above and beyond this initial reduction.

The second structure is the hierarchical polynomial relations described in Section 5.2. Recall that the order of a polynomial interaction is the total exponent of all variables in the interaction. Grouped variables like F will be treated as first order terms, so for example E^2F is a third order term. Given all lower order terms, terms of a given order will be assumed independent (the conditional independence principle). It is also assumed that the importance of a term X will depend only on terms of the next lowest order from which X inherits (inheritance and immediate inheritance principles).

The impact of these hierarchical relations is if a certain term is inactive, many other terms are unlikely (or forced) to be inactive. This property is explained by a diagram of relations among all terms, given in Figure 7.1. In each figure, terms are vertically grouped by order, with first order terms at the bottom. Connecting lines are not used to define individual relations because some terms have many parents and children. Instead superscripts and subscripts are used to indicate the number of children and parents for each term. The left half of Figure 7.1 gives counts for immediate parents and children, while the right half gives counts for the "extended family" of all parents and children (e.g. children, grandchildren, great-grandchildren, and so on). So the second order terms $[EH]_2^2$ (in the left of Figure 7.1) and $[EH]_2^5$ (in the right of Figure 7.1) indicate that EH has 2 immediate parents, 2 immediate children, and 2 parents and 5 children.

These figures make it possible to asses the impact of hierarchical structure on variable selection. For example, the left of Figure 7.1 shows that the first order terms E - H have more second order children than A - D. The right figure indicates that E, G, and H have 15 or 17 total children each, considerably more than the other first order terms. This means that if E, G, or H is inactive, a large number of higher order terms will be forced to be inactive (with strong heredity) or have reduced probability of being active (with weak heredity). These figures also make it possible to examine the dependence of specific high order terms on their parents. For example E^2G^2 has a total of seven parents while E^3G has six, implying that E^2G^2 has (slightly) less chance of being active. While the difference is minor here, in some cases more severe inequalities might occur.

In both of these examples, an examination of the number of parents and children of all terms in the model indicates which terms will have strong influence, and which are likely

Figure 7.1 Counts of immediate and extended parents and children (sub/superscripts), connector example.

Immediate Family	**Extended Family**
$[E^3G^2]_2^0 \quad [E^3H^2]_2^0$	$[E^3G^2]_{10}^0 \quad [E^3H^2]_{10}^0$
$[E^3F]_2^0 \quad [E^3G]_2^1 \quad [E^3H]_2^1 \quad [E^2G^2]_2^1$ $[E^2H^2]_2^1$	$[E^3F]_6^0 \quad [E^3G]_6^1 \quad [E^3H]_6^1 \quad [E^2G^2]_7^1$ $[E^2H^2]_7^1$
$[AG^2]_2^0 \quad [AH^2]_2^0 \quad [BG^2]_2^0 \quad [BH^2]_2^0 \quad [CG^2]_2^0$ $[CH^2]_2^0 \quad [DG^2]_2^0 \quad [DH^2]_2^0 \quad [EG^2]_2^1 \quad [EH^2]_2^1$ $[E^2F]_2^1 \quad [E^2G]_2^2 \quad [E^2H]_2^2 \quad [E^3]_1^3$	$[AG^2]_4^0 \quad [AH^2]_4^0 \quad [BG^2]_4^0 \quad [BH^2]_4^0 \quad [CG^2]_4^0$ $[CH^2]_4^0 \quad [DG^2]_4^0 \quad [DH^2]_4^0 \quad [EG^2]_4^2 \quad [EH^2]_4^2$ $[E^2F]_4^1 \quad [E^2G]_4^3 \quad [E^2H]_4^3 \quad [E^3]_2^5$
$[AF]_2^0 \quad [AG]_2^1 \quad [AH]_2^1 \quad [BF]_2^0 \quad [BG]_2^1$ $[BH]_2^1 \quad [CF]_2^0 \quad [CG]_2^1 \quad [CH]_2^1 \quad [DF]_2^0$ $[DG]_2^1 \quad [DH]_2^1 \quad [EF]_2^1 \quad [EG]_2^2 \quad [EH]_2^2$ $[E^2]_1^4 \quad [G^2]_1^5 \quad [H^2]_1^5$	$[AF]_2^0 \quad [AG]_2^1 \quad [AH]_2^1 \quad [BF]_2^0 \quad [BG]_2^1$ $[BH]_2^1 \quad [CF]_2^0 \quad [CG]_2^1 \quad [CH]_2^1 \quad [DF]_2^0$ $[DG]_2^1 \quad [DH]_2^1 \quad [EF]_2^2 \quad [EG]_2^5 \quad [EH]_2^5$ $[E^2]_1^{11} \quad [G^2]_1^7 \quad [H^2]_1^7$
$[A]_0^3 \quad [B]_0^3 \quad [C]_0^3 \quad [D]_0^3 \quad [E]_0^4 \quad [F]_0^5 \quad [G]_0^6$ $[H]_0^6$	$[A]_0^5 \quad [B]_0^5 \quad [C]_0^5 \quad [D]_0^5 \quad [E]_0^{17} \quad [F]_0^7 \quad [G]_0^{15}$ $[H]_0^{15}$

to be influenced by others.

7.2.3 Exploring Probabilistic Aspects of Priors

Graphics for assessing high dimensional structure, low-dimensional projections, and concentration of prior mass on specific models are introduced in this section. In practice they would be used primarily as a "reality check" to ensure that the priors used are compatible with expert knowledge. By using several graphics to look at the same high dimensional prior, different aspects of the prior structure should be revealed and made available for reality checking. Since no expert knowledge for the connector example is available, these plots will be introduced and used to differentiate the four priors under consideration. The section is not intended to catalog all possible graphics; rather it suggests some new plots that may be useful, and attempts to convey the exploratory nature of this process.

Sampling-Based Approaches

Calculations based on prior samples are well suited to the exploratory nature of this analysis. New plots may need to be rapidly prototyped, necessitating new probability calculations. Analytic results are usually more difficult than calculations based on prior samples, so the strategy will be to use prior samples wherever possible. Since the posterior analysis *must* be based on samples, using the same methods for priors saves time, reuses code, and treats the prior and posterior analyses symmetrically.

The analysis programs need little or no modification. The generation algorithms are simply a subset of the Gibbs sampling code, with no data and only the sampling of the $\boldsymbol{\delta}$s carried out. Direct sampling from the priors would be more efficient, but the availability of Gibbs code for the posterior makes Markov chain sampling attractive. Efficiency gains from direct sampling would be minimal, since posterior sampling will take much longer than either prior method. The posterior takes longer because of the additional parameters involved, which will slow convergence and increase computational effort for each Gibbs cycle.

Algorithms have been coded for both the independent (7.1), (7.2) and dependent (7.3), (7.4) priors. Samples from the independent priors may be generated with either algorithm, since independence is a special case of the dependent relations. This allows verification of the consistency of the (more complex) dependent algorithm. In fact, the samples are numerically identical as well as statistically indistinguishable, provided that the same starting values and random number seeds are used for both cases.

Exploring Low Dimensional Projections

The prior distributions for $\boldsymbol{\delta}$ are multidimensional, so one way to examine structure is through lower dimensional projections of the form $Q = \boldsymbol{\delta} R$. Here, $\boldsymbol{\delta}$ has dimensions $1 \times p$ and the projection matrix R is $p \times q$. Typically $q = 1, 2$, or perhaps 3, so as to make graphic display of the distribution of Q possible. Since samples are used to explore the prior on $\boldsymbol{\delta}$, the projection of the same samples will be used to explore the prior distribution of various Q's.

The $p \times 1$ projection $R = (1, 1, \ldots, 1)'$ has the most immediate interpretation - namely the number of active terms in the model, since $\boldsymbol{\delta}(1, \ldots, 1)' = \sum_{i=1}^{p} \delta_i$. Other interesting $(1 \times p)$ choices of R would include forms that count certain terms, for example

$$R_i = \begin{cases} 1 & \text{if } \delta_i \text{ is of order 2} \\ 0 & \text{otherwise} \end{cases} . \tag{7.5}$$

Other specific groups of variables could also be singled out, such as control factors, noise factors, or control by noise interactions. The distribution of these projections can be either considered marginally or in groups, although even three dimensions will be difficult to visualize.

Most of the projections of interest have weights of either zero or one, rather than a range of values of either signs. Negative values of R could be used to contrast the number of active terms in different groups, but are less useful. One case where R values other than 0/1 make sense is the connector experiment, which has different numbers of elements in $\boldsymbol{\delta}$ for different priors. For independence priors (7.1) and (7.2), there are 55 elements; for dependence priors (7.3) and (7.4) there are only 47, because terms involving the contrasts F_1 and F_2 share common $\boldsymbol{\delta}$'s. To correct this, one could count those $\boldsymbol{\delta}$'s twice by making $R_F = R_{AF} = \ldots = R_{E^3F} = 2$. However, since the difference in the number of terms is small, comparisons will simply be made on an unadjusted basis.

Figure 7.2 shows histograms of the four priors for several interesting projections. The first row corresponds to first order, the fifth to fifth order, and the sixth row to the total number of all terms. Several interesting features are noticeable from these plots:

- The independent priors (7.1) and (7.2) have similar distributions of the total number of active terms. This is probably attributable to the shared independence structure and the nearly equal expected values for the total number of terms. The dependent priors (7.3) and (7.4) differ from the independence priors, with more mass on the null model, and greater skew towards high values. This skew is caused by the requirements that lower order terms be active whenever higher order terms are active.

- Dependent priors (7.3) and (7.4) have identical distributions for the number of main effects. For counts of higher order terms, weak heredity (7.3) allows more terms to be active, since higher order terms need only one active parent.

- Independence priors (7.1) and (7.2) have different count distributions for individual orders, a fact attributable to the different marginals used in (7.2).

- Although it is tempting to visually construct the sixth row by adding together the other five, this only works for the independence priors. The correlations in the structural priors makes this "intuitive" approach incorrect.

The joint distributions of these projections may also be examined, although the display of discrete densities in more than two dimensions is difficult. The displays for the independent cases would be uninteresting, since there will be no correlations. For dependent priors (7.3) and (7.4), the plots illustrate considerable correlation between the counts for lower and higher order terms.

Assessing Probability Concentration on Models

Methods for examining the distribution of mass among models are another informative way of looking at the priors and the posteriors. The information that they present is high dimensional in nature, and complements the projections just considered. In this section, a plot that assesses the concentration of prior mass on models is introduced. Priors that concentrate mass on a small subset of the entire model space are desirable because they (hopefully) focus mass on the most interesting subset of the model space.

This concentration of prior mass may be thought of as a flexible way of approaching the problem of multiple testing. Although there may still be 2^p possible models to entertain, the focusing of prior mass on a subset of them makes the vast majority of these models highly improbable, and restrict the "data snooping" tendencies of automated variable selection procedures. The advantage that this Bayesian approach has is that it can downweight models without excluding them entirely. If the evidence in the data is strong enough for a model that is a priori unlikely, then the data will dominate the posterior, and that model will have a concentration of posterior mass.

Figure 7.2 Comparison of priors for the number of active model terms

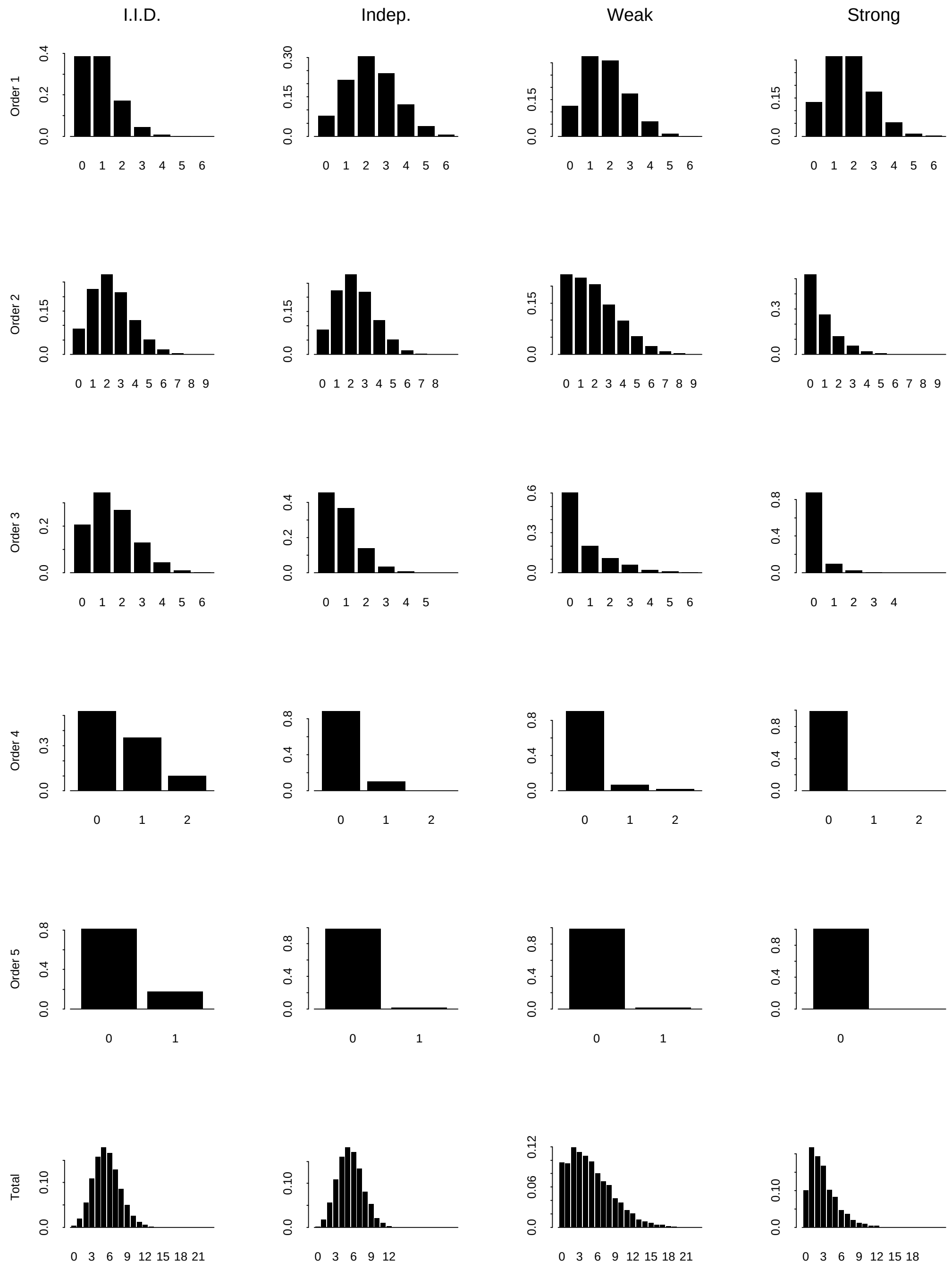
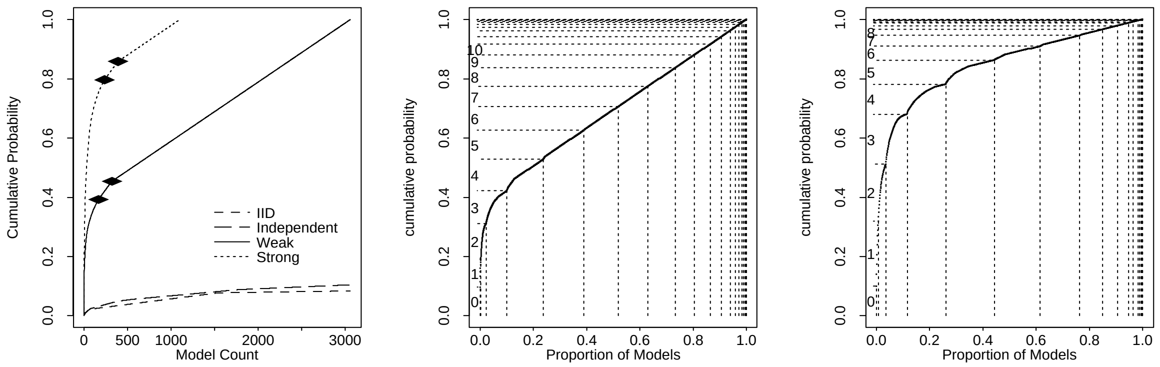


Figure 7.3 Concentration of prior mass on models, represented by cumulative model plots



The prior concentration of mass is an important quantity then, as it relates to the degree of prejudice towards (or against) subsets of the model space. A new plot is introduced to answers the question "How much mass is placed on the n most probable models?". The way it answers this is to plot model number n on the horizontal axis against cumulative probability for models $1, \ldots, n$ on the vertical axis. More probable models are placed to the left of the horizontal axis. That is, the y-coordinate for $x = 10$ would be the total probability mass of the ten models with the most mass. The line representing cumulative probability will start at zero on the left of the plot, and rise to the value one at the right hand side of the plot. These plots will be referred to as *cumulative model plots* because of their representation of cumulative probability across models.

The left plot in Figure 7.3 shows cumulative model plots for priors (7.1) through (7.4), with the lines for (7.1) and (7.2) calculated exactly, and lines for (7.3) and (7.4) based on samples. The ordering of the models along the horizontal axis is not the same for the four priors - the four lines are included on the same plot for the sake of comparison. The diamonds represent points where the mass associated with models drops to 2×5000^{-1} (the left diamond) and 1×5000^{-1} (the right diamond). Models to the right of these diamonds are visited so seldom that the corresponding probability estimates will be quite poor. Of course the true function will go towards one more gradually, but this bias is not worrying because the interesting models are the popular ones on left side of the plot.

This plot indicates that strong inheritance concentrates much more mass on a small subset of the model space. For example, (7.1), (7.2), (7.3) and (7.4) have prior probability masses of 0.023, 0.026, 0.351, and 0.687 to the first 100 models respectively. As will be shown in the posterior analysis, posteriors will deliver stronger conclusions when concentrated priors are used. A trade-off must be made between concentrating mass on interesting models and excluding (or assigning minimal mass to) models which might describe the data well.

The second and third plots in Figure 7.3 illustrated a variant of the cumulative model plot. For these plots, models are first grouped by the number of active terms, and then within each group, models are ordered by decreasing probability. Curvature of the line within a group indicates more concentration of prior mass on certain models in the group. For weak heredity (center plot), there is little concentration within groups of models that have more than four terms active. For strong heredity (right plot), models with up to five terms have differential concentration within groups.

The cumulative model plots illustrate that considerable mass is concentrated on a few models. "Which models?" is the next question. While these models could be examined as a matrix of zeros and ones, a related graphic display makes this exploration simpler. The 100×47 matrix of the 100 most popular models can be plotted as a square image with 47 horizontal values, and 100 vertical. Each element of the matrix will be plotted as either black (corresponding to a one) or grey (corresponding to a zero). By placing the matrix in a compact display, patterns can easily be recognized. To enhance the plot, labels of the terms are added along the horizontal axes, and a barplot of probability associated with each model (i.e. row) is placed to the right of the matrix plot. Note that the barplot is of individual and not cumulative probabilities.

A matrix plot for the 100 most popular models of prior (7.4) is given in Figure 7.4. Quite a few things can be learned from a close inspection of this figure. The null model is most popular, with a mass of around 0.10. After it, models with only one active main effect are the next most popular. It is interesting to note that not all models with only one main effect have the same mass. It appears that models with only one of A, B, C, D or F have approximately the same mass, and there is a drop in mass for models with only one of E, G or H active. A quick probability calculation reveals that this is true, since

$$\begin{aligned}\Pr(\delta_A = 1, \boldsymbol{\delta}_{(-A)} = 0) &= \Pr(\delta_A = 1)\Pr(\delta_B = 0)\ldots\Pr(\delta_H = 0)\\ &= 0.25 \times 0.75^7 = 0.0334\end{aligned} \tag{7.6}$$

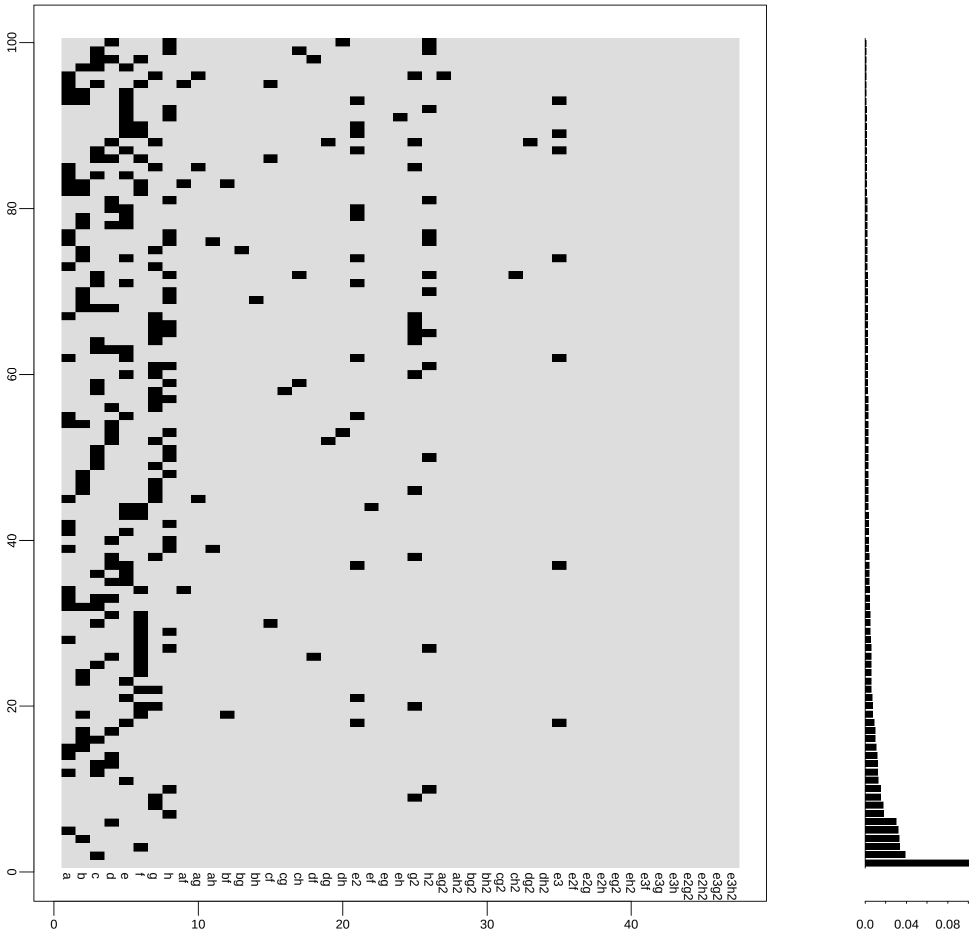
and

$$\begin{aligned}\Pr(\delta_H = 1, \boldsymbol{\delta}_{(-H)} = 0) &= \Pr(\delta_A = 0)\ldots\Pr(\delta_G = 0)\Pr(\delta_H = 1)\\ &\quad \times \Pr(\delta_{H^2} = 0|\delta_H = 1)\\ &= 0.25 \times 0.75^7 \times 0.5 = 0.0167.\end{aligned} \tag{7.7}$$

In (7.6), all other elements of the joint density are one, since $\Pr(\delta_{AB} = 0|\delta_A = 1, \delta_B = 0) = \Pr(\delta_{BC} = 0|\delta_B = 0, \delta_C = 0) = 1$. The same occurs in (7.7), except for the term $\Pr(\delta_{H^2} = 0|\delta_H = 1)$. This term differs because H is an active single parent. So although it was not immediately obvious, different prior masses have been assigned to models with single terms active when those terms are "single parents". What is also interesting is that the models $\{\delta_E = \delta_{E^2} = 1, \boldsymbol{\delta}_{(-E,-E^2)} = \mathbf{0}\}$ and $\{\delta_E = 1, \boldsymbol{\delta}_{(-E)} = \mathbf{0}\}$ have the same mass (the same goes for G and H). The sampling variation has separated the E, E^2 model from the other five models of this form although all six have exactly the same probability.

A more careful examination of the prior probabilities would have pointed this out, but this plot helps identify which probabilities to investigate, since there are many possible patterns to explore.

Figure 7.4 Which models are popular and their individual probabilities



7.2.4 Recap

What conclusions may be drawn from this analysis of priors? The structural analysis indicates that the grouping principle will reduce the model space, and that heredity acting on the large number of high order interactions will greatly reduce the chances that they are active. Projections indicate that independent priors favor models with a narrower range of active terms than dependent priors, whose distribution has both more mass on the null model and skew towards models with more terms. The cumulative model plots indicate that dependent priors (especially strong heredity) concentrate much more mass on a small region of the model space. The matrix plot of the priors allows examination of individual models, and identifies some interesting consequences of the choice of prior - namely the fact that some simple models assumed to be comparable have different prior probabilities.

7.3 Analysis of Posteriors for Variable Selection

Many of the analysis techniques presented in this section will be similar to those used to examine priors in Section 7.2. The same sampling based approach will be used, storing every tenth sampled value of $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta})$ from a single long run of length 50,000. The emphasis here will be on the probability content, not the structural aspects of the posterior, since the structure remains the same as specified in the priors. As with the prior analysis, the posteriors for $\boldsymbol{\delta}$ will be examined in different ways. Now the primary goal is to squeeze information from the posteriors, rather than reality checking.

Although not strictly correct, the names associated with priors will be used to identify corresponding posteriors. Thus, the I.I.D. posterior originates from an I.I.D. prior (7.1), and so on. Independent posteriors will refer to the two posteriors arising from independent priors on $\boldsymbol{\delta}$, and dependent posteriors will refer to the posteriors arising from dependent priors.

7.3.1 Numeric Evaluation of the Posteriors

In this section, various posterior probabilities are calculated from the samples based on priors (7.1) to (7.4). These numeric summaries of the posteriors are provided before more detailed plots are examined because reference will be made to these numeric summaries later. Table 7.3 lists the most probable models for each of the four posteriors. The term F without a subscript refers to both factors F_1 and F_2, and is used to emphasize the effect grouping utilized in priors (7.3) and (7.4). Only three models are given for the I.I.D case because all other models were only visited once (and would have estimated probability N^{-1} for a posterior sample of size N, here 5000). The difference in concentration of posterior mass is striking, especially for I.I.D. prior (7.1). Another model of interest, $\boldsymbol{\delta} = \mathbf{0}$, the model with no active factors was never visited by any of the four procedures, and consequently receives zero mass. A more detailed analysis of the concentration of posterior mass on models is presented later.

Table 7.3 The most probable models for the four posteriors

I.I.D.		Independent	
$A,D,E^2,F_1,F_2,G,G^2,H,$		A,D,F_1,F_2,G	0.0412
$BG^2,CF_1,DG^2,DH,$		A,D,F_1,G	0.0364
DH^2,EG,EH,E^3F_1	0.0004	A,D,F_1,F_2,G,DH	0.0182
A,D,F_1,F_2,G	0.0004	A,D,F_1,G,DH	0.0110
A,D,F_1,G,BG^2,DH	0.0004	A,D,E^2,F_1,F_2,G	0.0098
		A,D,F_1,F_2,G,BG	0.0076
		A,D,E^2,F_1,F_2,G,DH	0.0070
		A,D,E^2,F_1,G	0.0064
		A,D,F_1,F_2,G,CF_1	0.0060
		A,D,F_1,F_2,G,H	0.0048

Weak Heredity		Strong Heredity	
A,D,F,G	0.0470	A,D,F,G	0.1508
A,D,F,G,DH	0.0262	A,D,F,G,G^2	0.0756
A,D,F,G,BG,DH	0.0138	A,D,F,G,DG	0.0528
A,D,F,G,DG	0.0134	A,D,F,G,AG	0.0386
A,D,F,G,BG	0.0130	A,D,F,G,DF	0.0332
A,D,F,G,DH,G^2	0.0122	A,D,F,G,AF	0.0314
A,D,F,G,DF	0.0108	A,D,F,G,DF,DG	0.0214
A,D,F,G,AG	0.0104	A,D,F,G,AF,G^2	0.0188
A,D,G,DH	0.0104	A,D,F,G,DF,G^2	0.0184
A,D,F,G,G^2	0.0090	D,F,G	0.0176

The marginal probability that each term is active is also of interest, and is presented in Table 7.4. Only those terms with marginal posterior probability of 0.25 or more are included. One thing is clear from these tables - there appears to be considerable agreement (from marginal probabilities at least) that the terms A, D, F, G are active. Several other second and higher order terms might also be significant, but there is less agreement on this. Since the A, D, F, G model seems possible, why not calculate the probability of this event? Table 7.5 gives this probability for the four posteriors, calculated by averaging over all other elements of $\boldsymbol{\delta}$. While the probability of this event varies considerably across the four models, there is evidence from each model that these four terms are important. The small probability for the independent case is due to the small marginal probability that F_2 is active (in Table 7.4). One possible reason for this is that the i.i.d. model makes it easier for higher order terms to be active, and so it prefers them to F_2.

The fact that this is a robust design experiment will also influence the terms that are included in the model. Even without refitting the model with the robust design priors suggested in Section 6.1, several comments may be made. First, the control by noise interaction DH has considerable mass under all but the strong heredity prior. Control by noise

Table 7.4 Marginal probabilities for individual terms

I.I.D.		Independent	
G	1	G	1
D	0.996	D	0.991
A	0.996	A	0.967
F_1	0.96	F_1	0.912
F_2	0.829	F_2	0.648
DH	0.748	DH	0.343
E^2	0.6		
E^3G^2	0.45		
E^3F_2	0.433		
CF_1	0.383		
BG	0.339		

Weak Heredity		Strong Heredity	
G	0.999	G	0.999
D	0.995	D	0.982
A	0.954	F	0.938
F	0.943	A	0.93
DH	0.577	G^2	0.356
BG	0.336	DG	0.274
G^2	0.324		
DG	0.267		

Table 7.5 Marginal probability that the most likely terms are all active

$\Pr(A, D, F_1, F_2, G \text{ active})$	
I.I.D.	0.795
Independent	0.582
Weak Heredity	0.902
Strong Heredity	0.866

interactions mean an opportunity to improve the process, so the final model used should include this factor if the goal is robust design optimization. Other terms might also be included for further study, such as BG, DG, or G^2. All these involve noise factors, so again they indicate a possible opportunity for improvement.

7.3.2 Decomposing Group Effects

All the posteriors give fair to strong evidence that F_1 and F_2 are important predictors. Posteriors (7.3) and (7.4) make a blanket statement about both factors - namely "It looks like heating (i.e. factor F) makes a difference". The next question to ask is: "If heating makes a difference in the response, is one of the heating methods accounting for this difference?".

More generally, a meaningful decomposition of the overall group effect is desired which will split the effect into components that influence the response, and remainder components which do not. This is not done at the onset because it is not certain a priori that *as a group*, the terms are important. For example, suppose six heating methods were used and none of them really made a difference. If grouping is not used in the analysis, the chances are large that one of them will appear influential because of random variation in the response. By grouping the terms together, the chance of identifying spurious effects is reduced. Once a group is identified as influential, it may be decomposed into meaningful components if any are evident.

Since this group consists of only two coefficients, decompositions can be identified by examining a scatterplot of the posterior sample, shown in Figure 7.5. Note that the coefficients have been reparameterized here, so that they represent contrasts between the ambient level and the other two settings. The range of the marginals implies that 15@350 has an effect larger than ambient, and 10@400 has an effect smaller than ambient. The correlation pattern implies that if there is a large difference between ambient and 10@400, then the difference between ambient and 15@350 will be in the opposite direction.

This decomposition is not very informative; in many cases simplifications will be possible. For example, if the two heating treatments were identical and different from the ambient method, the posterior would concentrated around a point distant from the origin. This would be a more meaningful decomposition because the three distinct levels of the heat treatment are put into two groups that make a difference.

7.3.3 Exploring Low Dimensional Projections

The same six projections that were explored for the priors will be again examined for the posteriors. Recall that the largest differences were between the independent and structured priors. The same six projections of the posteriors resulting from priors (7.1) to (7.4) are given in Figure 7.6, and the posterior expected number of active terms is presented in Table 7.6. There is considerable agreement among three of the four posteriors - namely all but the i.i.d. one. In general, the posteriors resulting from the i.i.d. prior tend to favor more active terms in the model, especially in the case of higher order terms. Since prior (7.1) places by far the most mass on active terms of orders 3-5, this is not surprising.

Figure 7.5 Posteriors of the coefficients for two grouped contrasts.

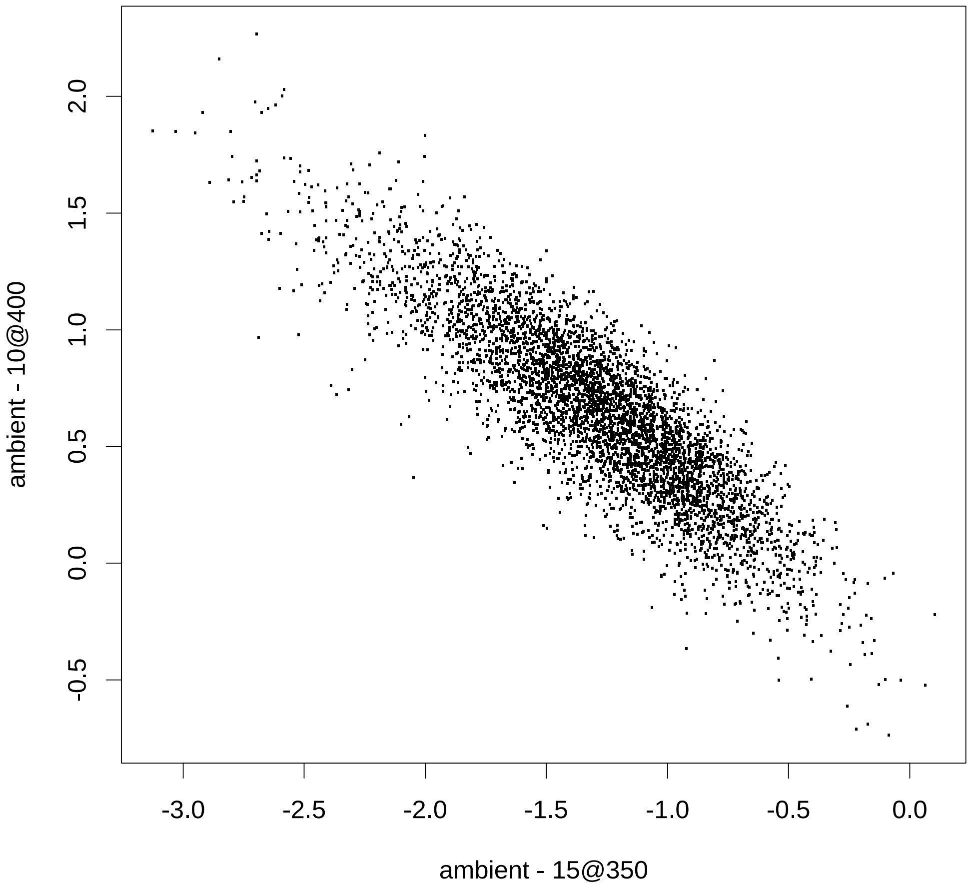


Figure 7.6 Comparison of posteriors for the number of active model terms

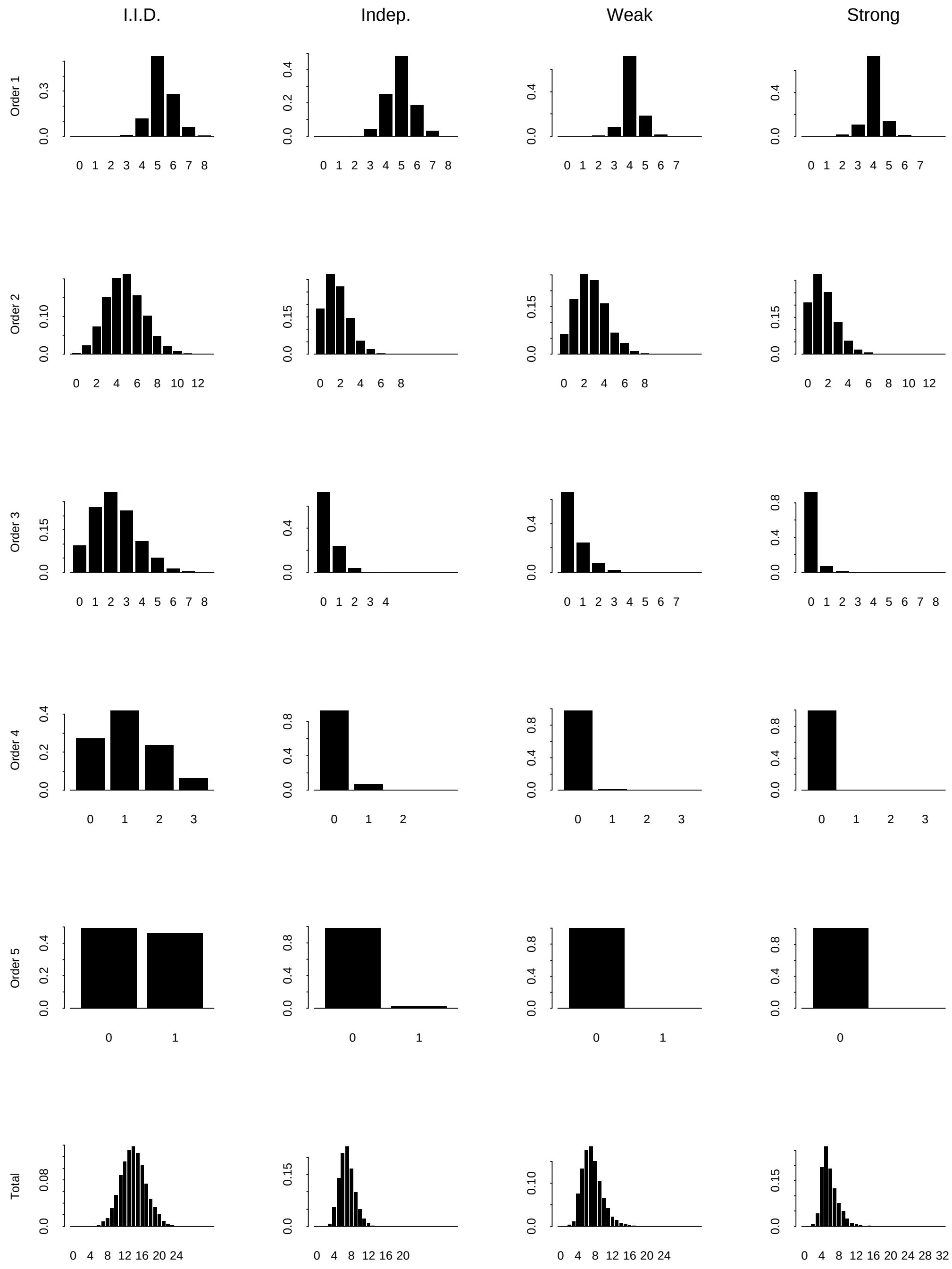


Table 7.6 Expected number of active terms, grouped by order

	Order					
Prior	1	2	3	4	5	All
I.I.D.	5.28	4.82	2.24	1.12	0.56	14.01
Indep.	4.92	1.65	0.32	0.07	0.02	6.99
Weak	4.11	2.67	0.47	0.03	0.00	7.35
Strong	4.03	1.60	0.09	0.00	0.00	5.90

It is somewhat surprising that the posteriors for the other three priors agree so closely. There are of course minor differences (for example less certainty in the number of main effects with an independence prior), but on the whole, the patterns are quite similar. This implies that the data overwhelm many differences among priors along these projected directions.

7.3.4 Examining the Multivariate Structure of the Posteriors

In this section, matrix plots will be used to explore the posteriors in a slightly different way than they were used for the priors. With the priors, many variables were treated as "equals", with interchangeable probability structures. For example, factors A through D have the same probability structure, and the same types and numbers of children. The same is not true for the posteriors, where the data has added information about which factors are important. Now it is important to see how the data have modified the structures specified in the prior and what relationships between predictors are suggested. This is of course a multivariate undertaking, and one for which the matrix plots are well suited.

The specific version of the matrix plot used here will have each row as a single observed value of $\boldsymbol{\delta}$, as before. The ordering of the rows will simply be as they are sampled, rather than in decreasing order of occurrence or any other scheme. The idea is to provide a view of the samples as they are drawn, including any repetitions, if they occur. By not ordering the vertical axis, this plot serves as both an exploratory and diagnostic tool. If the algorithm is not converging or exhibiting unusual behavior, it may turn up in the matrix plot.

The matrix plots for the four posteriors are given in Figures 7.7 to 7.10. Because of the limited resolution of the plot, only the last 500 samples (4500-5000) are displayed. Diagnostic use of samples 1-500 revealed no different patterns, so the last samples are used here.

Quite a few interesting things may be said about these plots (recall that the posteriors are identified by the names assigned the corresponding priors):

- All four indicate that the terms A, D, F_1, F_2, and G appear to be active, a fact supported by the numeric summaries given previously (F_1 and F_2 are represented by F for posteriors (7.3) and (7.4)).

- The ordered nature of the plot (i.e. as sampled) indicates that the posteriors have a

Figure 7.7 Posterior matrix plot for connector data: I.I.D. prior (7.1)

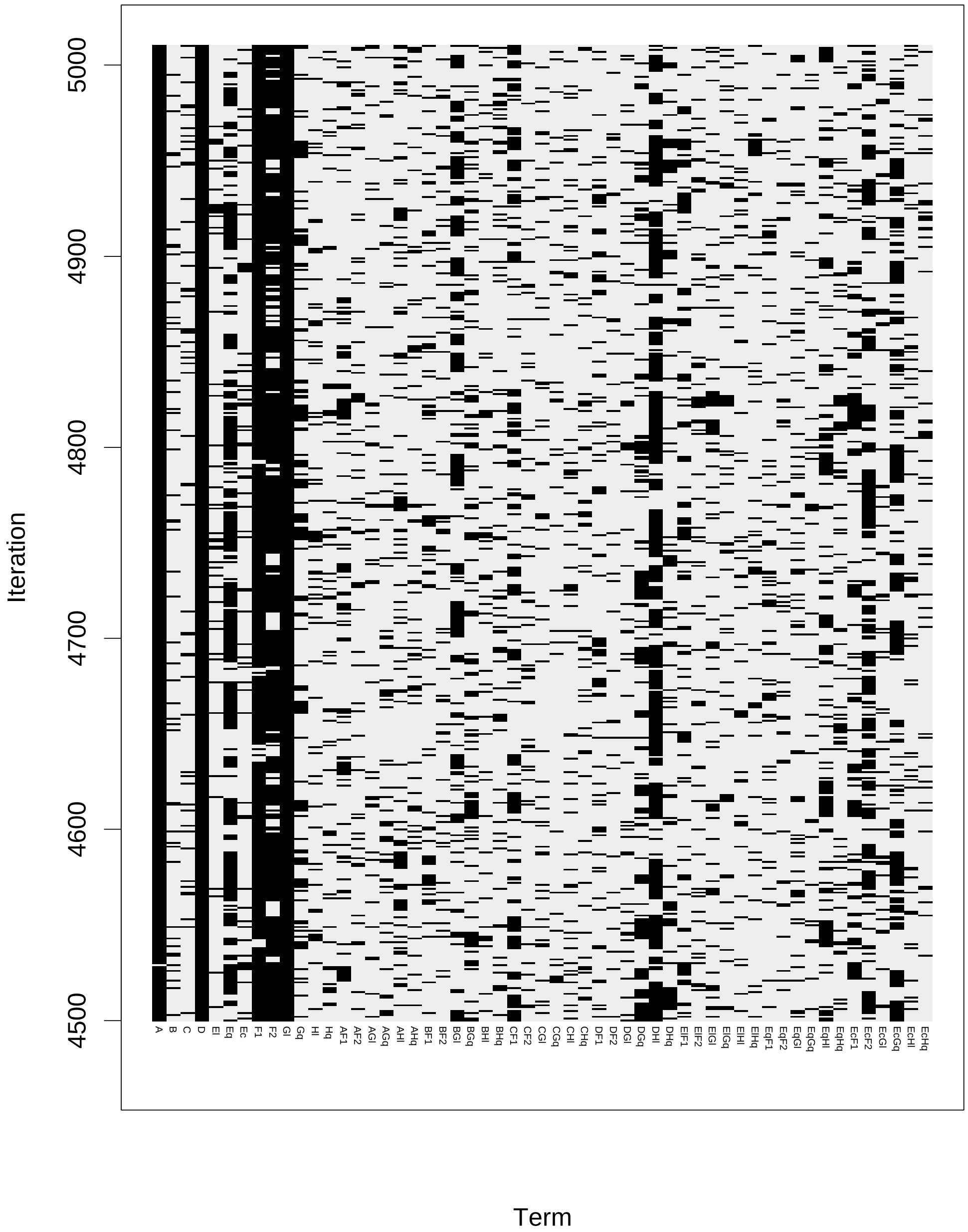


Figure 7.8 Posterior matrix plot for connector data: independent prior (7.2)

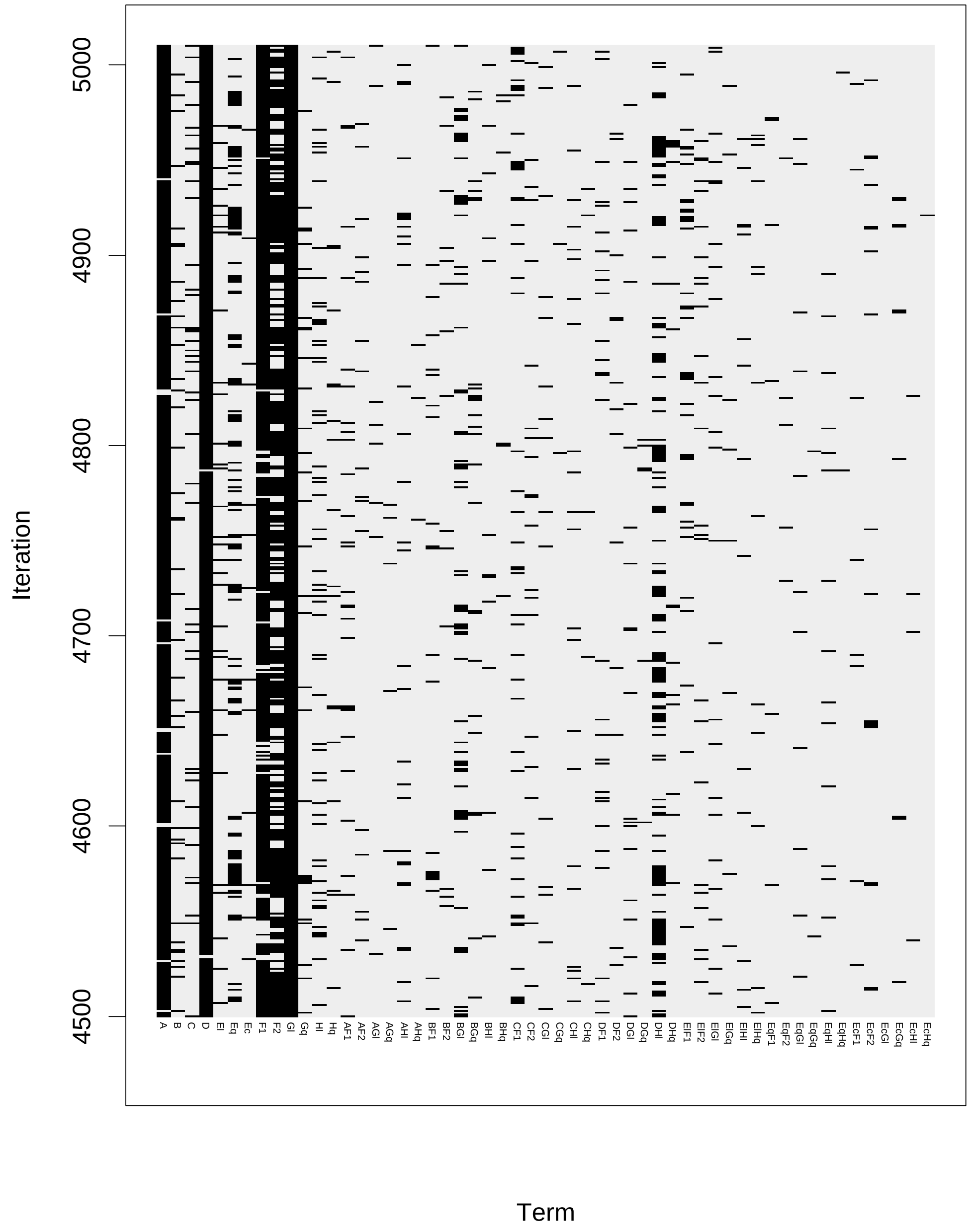


Figure 7.9 Posterior matrix plot for connector data: weak heredity prior (7.3)

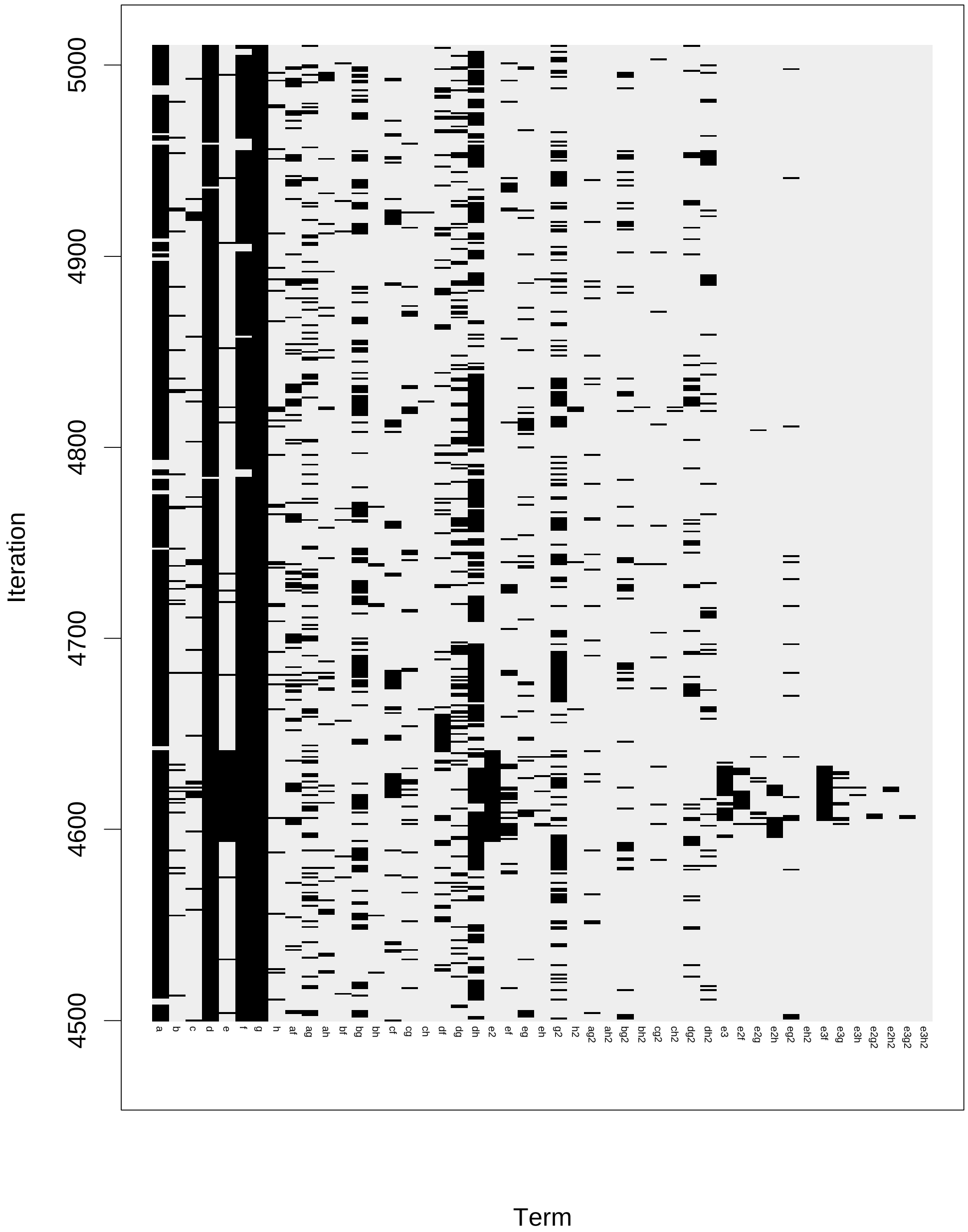
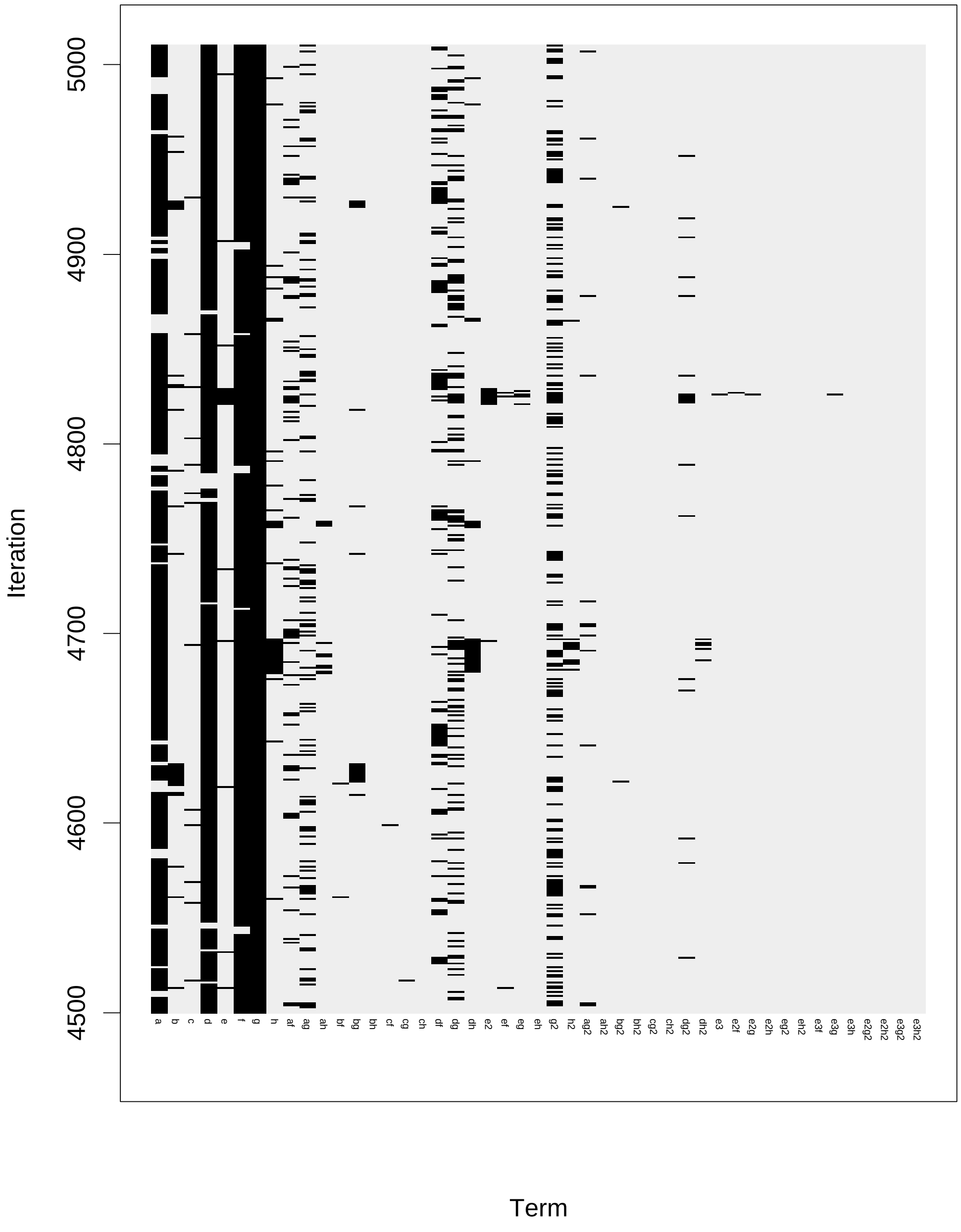


Figure 7.10 Posterior matrix plot for connector data: strong heredity prior (7.4)



tendency to contain long runs in which a factor is either active or inactive, but not changing state. This is due to the high-dimensional parameter space, which the Gibbs sampler traverses in a very conditional way (especially with the elements of $\boldsymbol{\delta}$) This aspect of the plots suggests that to attain valid inference, long runs should be used, and not every observation taken, a strategy that was employed to draw samples for this analysis. For further discussion of convergence issues related to the $\boldsymbol{\delta}$'s and in general, see Section 5.6.1.

- Independent posteriors seem to have less structured patterns in the higher order terms than dependent posteriors, probably because of the independence structure. Both independent posteriors (Figures 7.7 and 7.8) seem to present the same information, but the I.I.D posterior has more noise. This additional noise is because the prior makes it easier for higher order terms to enter the model.

- The weak heredity posterior (Figure 7.9) has a considerable number of cases in which higher order terms are active. There are in fact more active second order terms than the independence posterior (Figure 7.8). It is interesting that this pattern occurs, since the marginals of the two priors are equal. Notice that although there are many 1's for interactions under weak heredity, they tend to be only for certain terms, namely those with at least one active parent. Although the marginal probability of an active interaction is small, conditional on an active parent, the chances of a term becoming active jumps to 20% (and 50% for a single parent term). Since some main effects appear to be active, their children are more likely to enter the model. So the independence prior is more likely to spread out the probability of active higher order terms, and heredity priors will take that same mass and concentrate it on a few terms.

- The strong heredity posterior (Figure 7.10) has far fewer terms active, because of the much more restrictive prior used. Higher order terms are scarce, since they depend on lower order terms. For example, both E and H appear to be inactive, so the strong heredity prior would say that conditional on these main effects being inactive, 27 terms of order two to five *must* be inactive.

- A close examination of Figure 7.10 (strong heredity) shows that certain terms are included almost all the time that their parents are. For example, the DH term is active in four different runs - before 4700, near 4750 and 4790, and 4860. Strong heredity means that H must be active in all of these samples as well. What is interesting is that it is not active at many other times. That is, DH seems to be in the model whenever it gets a chance. This suggests that the strong heredity restriction is limiting this term, and that if a less restrictive prior was used, then that term would be active more often. A quick glance at the weak heredity posterior (Figure 7.9) confirms this - the DH term is active much more than its H parent once strong heredity is removed.

For this data the plots seems in agreement, a fact attributable to the relatively small number of active terms suggested by the data. The differences between the plots are minimal, and any evidence for higher order terms is not strong.

7.3.5 Assessing Probability Concentration on Models

In this section, the *cumulative model plots* introduced in Section 7.2 are used as a measure of concentration of posterior mass. The numeric summaries presented in Section 7.3.1, indicate that i.i.d. posterior (7.1) visits only three models more than once, and so it is not considered in this section. Figure 7.11 contains plots for the posteriors (7.2) to (7.4). The left column of plots are unordered, and compare priors to posteriors. Only the first 500 models are plotted on the horizontal axis, since certainly no more than that many models would be of interest. The right column of plots have models grouped along the horizontal axis by the number of active terms.

The greatest increase in concentration from prior to posterior occurs from using the independent prior (7.2). The change in concentration for dependent priors (7.3) and (7.4) is much smaller, although the posterior concentration is still larger than with an independent prior. The vast differences in posterior concentration of mass between I.I.D. prior (7.1) (not shown) and independent prior (7.2) is attributable to the fact that prior (7.1) favors all terms equally, while (7.2) places a reduced mass on 27 of the 55 terms. By concentrating more mass on fewer models, the independence prior makes greater posterior concentrations possible.

A first look at these plots would suggest that concentration has not changed much from prior to posterior for (7.3) and (7.4). Although the pattern of concentration does not change drastically, the models receiving the mass do change. For example, the null model receives the most prior mass, but is never visited in the posterior sample.

Another way of looking at the (say) 100 most popular models is to use the matrix plot. This is illustrated for the case of the 100 most popular models suggested by strong heredity posterior (7.4) in Figure 7.12. The numeric summaries have already indicated the first 10 most popular models. This plot gives a graphic overview of which terms occur most often in the popular models. Obviously, A, D, F, and G are common to almost all the popular models. Terms such as DF, DG, DH, and G^2 appear quite often as well. Other terms seldom appear, and quite a few terms never appear in the "top 100" models. This is attributable to the strong heredity principle, and the fact that only main effects A, D, F, and G appear active.

7.4 Conclusions

This analysis has revealed several things about the priors and posteriors. It is clear that the main effects A, D, F and G are active. Several other interactions and higher order terms might be active, and could be considered in the model. This is especially important since this experiment is a robust design experiment, which places added importance on control by noise interactions. Several terms involving noise among the borderline group, including

Figure 7.11 Posterior cumulative model plots under priors (7.2)-(7.4)

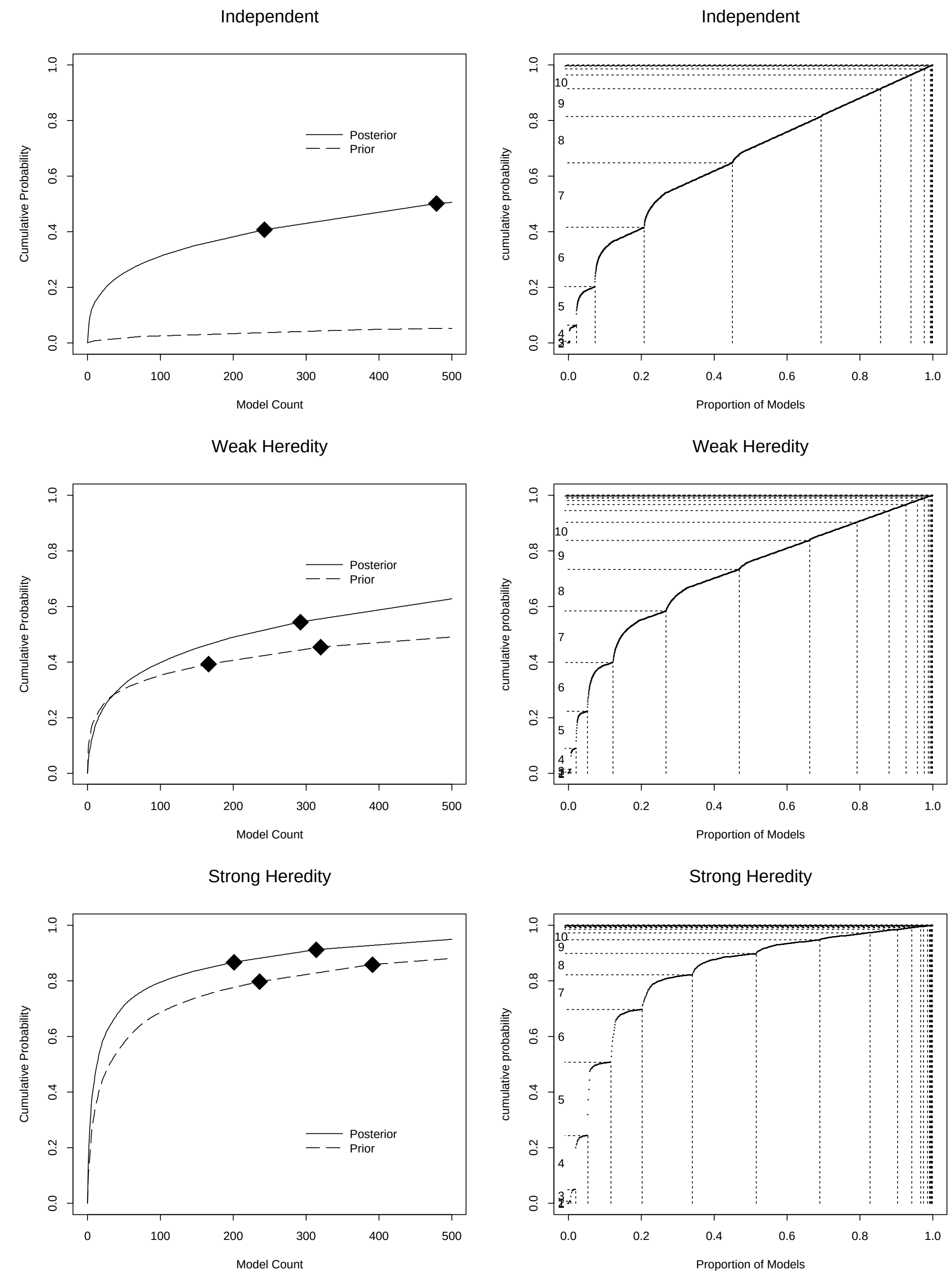
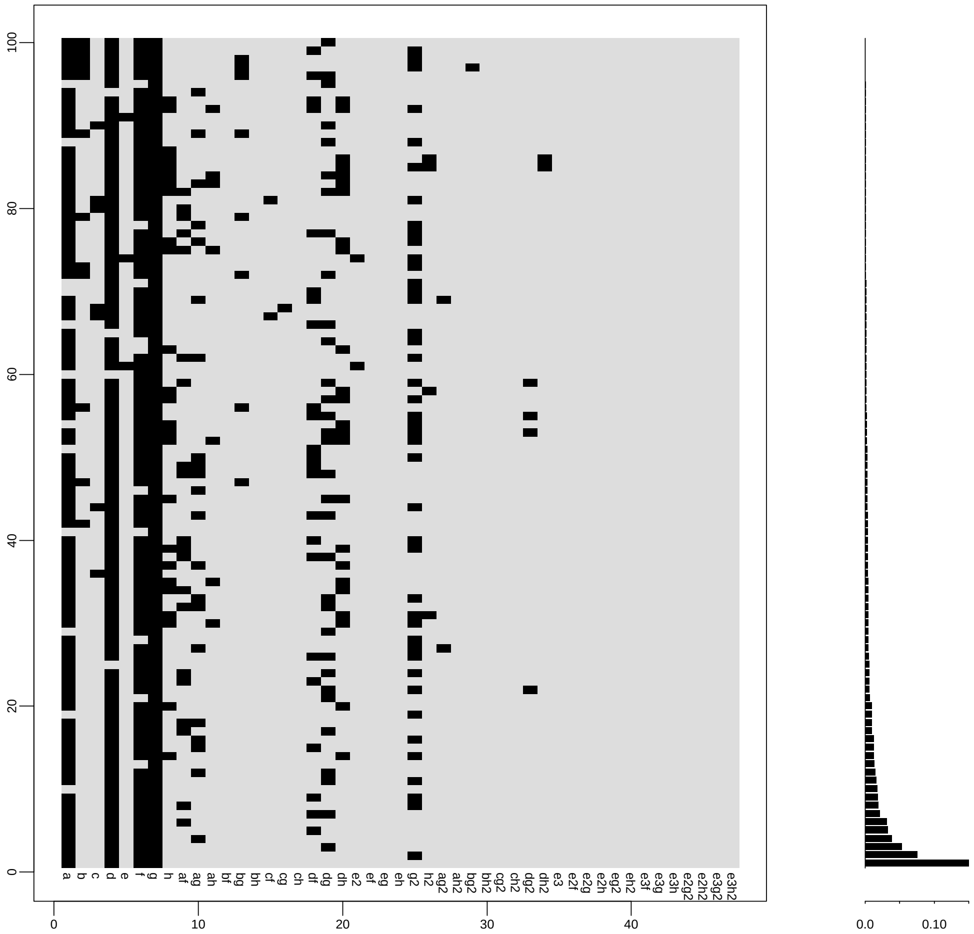


Figure 7.12 The 100 most popular models for strong heredity posterior



control by noise interactions DH, BG, DG, and noise effect G^2. Since these terms represent an opportunity to improve the process, they are likely to be included in the model.

For simple model identification, the most useful tools were tables of probabilities for full models and marginals, and the image plot of the 100 most popular models. The histograms of projected posteriors were similar for the four priors, and indicated that 5-10 terms were active. The matrix plots of $\boldsymbol{\delta}$ as it was sampled indicated considerable correlation between samples and made it possible to inspect the high dimensional structure of the posterior. The cumulative model plots indicated that even when priors concentrated considerable mass on a small region of the model space, no single model was clear cut in the posteriors.

The next step would be to refit a model with the terms that appear important (A, D, F, G, and possibly G^2 and other control by noise interactions), and then optimize the process using the procedures of Section 2.3.1. This will not be described here, as no new procedures are required.

Chapter 8

Summary

Data with an ordered categorical response occur quite commonly in industrial experiments. As an attempt to better model the relation between this response and predictor variables, McCullagh's GLM is considered. Although Bayesian versions of this model had been considered before, by Albert and Chib (1993) for example, the motivation here is new. A Bayesian approach with informative priors is used to resolve the problem of infinite estimates. A sampling based approach, the Gibbs sampler, was used to obtain posteriors for the models considered.

An additional advantage of the Bayesian approach is evident when the fitted model is used to optimize the process. Previous approaches typically assume that the fitted model and point estimates constitute "the truth", and optimize over this. The point estimates fail to capture information about the uncertainty in the parameters. The optimization under the Bayesian approach is over the posteriors, and consequently acknowledges this uncertainty, resulting in a more realistic optimization process. Optimization for robust design experiments is also considered, and response surface approach is easily and naturally integrated into the optimization over the uncertainty in posteriors.

The variable selection component of the procedure introduced in the first part of the thesis has several difficulties, including approximate criteria that do not account for prior information, no formal acknowledgment of dependence between some terms like interactions and main effects, and a non-automatic algorithm. An automated procedure that allows prior specification of the probability that a term will be active is reviewed, and generalized so that it will be applicable to industrial problems. The generalization is to introduce dependence in the prior probabilities that certain terms are active. These dependencies include relations between main effects and (possibly higher order) interactions, the ability to group terms, competing predictors, and global constraints on aspects of the model such as sample size. It remedies the shortcomings of the non-automated approach by utilizing priors that incorporate dependence among terms, using an automated algorithm and a model selection criterion that is exact and depends on the priors specified. The mathematical specification of prior beliefs (or "principles") about dependence is precise yet flexible. The automated algorithm makes it possible to consider the entire model space without refitting the model repeatedly. The criterion used to select models is the posterior probability of each model, a natural and non-approximate criterion applicable to small samples.

All relations introduced may be used to exclude certain models from the selection procedure. Heredity and global weighting relations offer additional flexibility. They can assign small amounts of mass to models that seem unlikely, instead of excluding them entirely. With the majority of prior mass focused on a small subset of the model space, the posteriors will be concentrated on models within this subset unless there is strong counter evidence. This flexibility is something not offered by frequentist procedures, and is an important advantage of this procedure.

Each of the four relations offers specific advantages. Interaction relations include the ability to specify whether an interaction needs one or all parents active to be included in the model. It also makes it possible to discriminate against higher order terms, which imply added complexity and reduced stability in prediction. The grouping principle makes it possible to determine whether several related factors are active as a group. This is especially useful when a parameterization is not unique, and variable selection for individual terms in the group would be parameterization dependent. After groups of terms have been identified as important, their effect can be decomposed into meaningful contrasts without refitting the model. Competing predictors allow external information about exclusive relations to be incorporated in the model, and can be used to identify transformations of predictor terms. Global constraints make it possible to reweight model probabilities by global properties, such as the number of active terms.

Although not fully demonstrated in the case study of Chapter 7, this method will be useful under any of the following circumstances: Models with considerable correlation structure within $\boldsymbol{\beta}$, models with rich relations between predictors, and models with as many or more predictors as data points.

The main advantage of this variable selection approach is the flexibility and speed it offers. In combination with the Bayesian GLM of Chapter 2, the methods in this thesis provide an integrated way to model ordinal data, and suggest possible applications to other industrial problems.

Bibliography

Aitkin, M. (1991), "Posterior Bayes Factors" (with discussion), *Journal of the Royal Statistical Society, Series B*, 53, 111-142.

Albert, J. and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Booth, K. H. V. and Cox, D. R. (1962), "Some Systematic Supersaturated Designs," *Technometrics*, 4, 489–495.

Box, G. E. P. (1980), "Sampling and Bayes' Inference in Scientific Modelling and Robustness" (with discussion), *Journal of the Royal Statistical Society, Series A*, 143, 383–430.

Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978), *Statistics for Experiments: An Introduction to Design, Data Analysis, and Model Building*, John Wiley & Sons, New York.

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

Box, G. E. P. and Meyer, R. D. (1993) "Finding the Active Factors in Fractionated Screening Experiments," *Journal of Quality Technology*, 25, 94–105.

Carlin, B. P., and Gelfand A. E. (1991), "An Iterative Monte Carlo Method for Nonconjugate Bayesian Analysis," *Statistics and Computing*, 1, 119–128.

Draper, N. R., and Smith, H. (1981) *Applied regression analysis (Second Edition)*, John Wiley & Sons, New York.

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.

Gelfand, A. E. (1992), Discussion of "Practical Markov Chain Monte Carlo," *Statistical Science*, 7, 486-487.

Gelman, A. and Rubin, D. B. (1992), "Inference from Iterative Simulation Using Multiple Sequences" (with discussion), *Statistical Science*, 7, 457–472.

Geman, S. and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

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

George, E. I. and McCulloch, R. E. (1994), "Fast Bayes Variable Selection," Technical Report, Graduate School of Business, University of Chicago.

George, E. I., McCulloch, R. E. and Tsay, R. S. (1994), "Two Approaches to Bayesian Model Selection with Applications," Technical Report, Graduate School of Business, University of Chicago.

Geyer, C. J. (1992), "Practical Markov Chain Monte Carlo" (with discussion), *Statistical Science*, 7, 473-483.

Gilks, W. R., and Wild, P. (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Applied Statistics*, 41, 337–348.

Hamada, M. and Wu, C. F. J. (1990), "A Critical Look at Accumulation Analysis and Related Methods", (with discussion), *Technometrics*, 32, 119–162.

Hamada, M. and Wu, C. F. J. (1992), "Analysis of Censored Data from Fractionated Experiments: a Bayesian Approach," The Institute for Improvement in Quality and Productivity Research Report RR-92-11, University of Waterloo.

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

Jeffreys, H. (1961), *Theory of Probability, Second Edition*, Clarendon Press, Oxford.

Jinks, J. (1987), "Reduction of Voids in a Urethane-Foam Product," in *Fifth Symposium on Taguchi Methods*, Dearborn, MI: American Suppliers Institute, Inc., 135–148.

Kernighan, B. W. (1977), "RATFOR - A Preprocessor for a Rational Fortran," Computing Science Technical Report 55, Bell Laboratories.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988), "Local Computations with Probabilities on Graphical Structures and their Application to Expert Systems" (with discussion), *Journal of the Royal Statistical Society, Series B*, 50, 157–194.

Lawless, J. F. and Singhal, K. (1978), "Efficient Screening of Nonnormal Regression Models," *Biometrics*, 34, 318–327.

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

Madigan, D. and Raftery, A. E. (1991), "Model Selection and Accounting for Model Uncertainty in Graphical Models using Occam's Window," Technical Report 213, Department of Statistics, University of Washington.

McCullagh, P. (1980), "Regression Models for Ordinal Data" (with discussion), *Journal of the Royal Statistical Society, Series B*, 42, 109–142.

McCullagh, P. (1988), "PLUM: An Interactive Package for Analysis of Ordinal Data," Technical Report, Department of Statistics, University of Chicago.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models (Second Edition)*, Chapman & Hall, London, UK.

Miller, A. J. (1990), *Subset Selection in Regression*, Chapman & Hall, London.

Mitchell, T. J. and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression" (with discussion), *Journal of the American Statistical Association*, 83, 1023–1032.

Nair, V. N. (1986), "Testing in Industrial Experiments with Ordered Categorical Data" (with discussion), *Technometrics*, 28, 283–291.

Peixoto, J. L. (1987), "Hierarchical Variable Selection in Polynomial Regression Models," *American Statistician*, 41, 311-313.

Peixoto, J. L. (1990), "A Property of Well-Formulated Polynomial Regression Models," *American Statistician*, 44, 26–30.

Quinlan, J. (1984), "Study to Determine Factors Affecting Solvent Bonded Connector Durability", in *Second Symposium on Taguchi Methods*, Dearborn, MI: American Suppliers Institute, Inc., 78–100.

SAS Institute, Inc. (1989), *SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 2*, SAS Institute, Inc., Cary, NC.

Schervish, M. J. and Carlin, B. P. (1992), "On the Convergence of Successive Substitution Sampling," *Journal of Computational and Graphical Statistics*, 1, 111–127.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–464.

Shoemaker, A. C., Tsui, K. L. and Wu, C. F. J. (1991), "Economical Experimentation Methods for Robust Design," *Technometrics*, 33, 415–427.

Smith, A. F. M. (1991), "Bayesian Computational Methods," *Phil. Trans. of Royal Soc. of London, Series A*, 337, 369–386.

Smith, A. F. M. and Spiegelhalter, D. J. (1980), "Bayes Factors and Choice Criteria for Linear Models," *Journal of the Royal Statistical Society, Series B*, 42, 213–220.

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

Spiegelhalter, D. J. and Smith, A. F. M. (1982), "Bayes Factors for Linear and Loglinear Models with Vague Prior Information," *Journal of the Royal Statistical Society, Series B*, 44, 377–387.

Statistical Sciences, Inc. (1991), *S-PLUS User's Manual*, Statistical Sciences, Inc., Seattle, WA.

Steinberg, D. M., and Bursztyn, D. (1993), "Confounded Dispersion Effects in Robust Design Experiments with Noise Factors," Center for Quality and Productivity Improvement Report 93, University of Wisconsin-Madison.

Taguchi, G. (1974), "A New Statistical Analysis for Clinical Data, the Accumulating Analysis, in Contrast with the Chi-Square Test," *Saishin Igaku*, 29, 806–813.

Tanner, M. and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation", *Journal of the American Statistical Association*, 82, 528–550.

Tierney, L. (1994), "Exploring Posterior Distributions Using Markov Chains," Technical Report No. 560 (Revised), School of Statistics, University of Minnesota.

Tierney, L. and Kadane, J. B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86.

Tse, S. K. (1986), "On the Existence and Uniqueness of Maximum Likelihood Estimates in Polytomous Response Models," *Journal of Statistical Planning and Inference*, 14, 269–273.

Wakefield J. C., Gelfand A. E., and Smith, A. F. M. (1991), "Efficient Generation of Random Variates via the Ratio-of-uniforms Method," *Statistics and Computing*, 1, 129–133.

Welch, W. J., Yu, T. K., Kang, S. M. and Sacks, J. (1990), "Computer Experiments for Quality Control by Parameter Design," *Journal of Quality Technology*, 22, 15–22.

Wilks, S. S. (1962), *Mathematical Statistics*, John Wiley & Sons, New York.

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

Zellner, A. and Rossi, P. E. (1984), "Bayesian Analysis of Dichotomous Quantal Response Models," *Journal of Econometrics*, 25, 365–393.