Bayesian Analysis of Ordered Categorical Data from Industrial Experiments

Hugh Chipman
Graduate School of Business
University of Chicago
Chicago, IL 60637

Michael Hamada
Department of Statistics and Actuarial Science and
The Institute for the Improvement of Quality and Productivity
University of Waterloo
Waterloo, Ontario, N2L 3G1, Canada

March, 1995

Abstract

Data from industrial experiments often involve an ordered categorical response, such as a qualitative rating. ANOVA based analyses may be inappropriate for such data, suggesting the use of Generalized Linear Models (GLMs). When the data are observed from a fractionated experiment, likelihood-based GLM estimates may be infinite, especially when factors have large effects. These difficulties are overcome with a Bayesian GLM, which is implemented via the Gibbs sampling algorithm. Techniques for modeling data and for subsequently using the identified model to optimize the process are outlined. An important advantage in the optimization stage is that uncertainty in the parameter estimates is accounted for in the model. For robust design experiments, the Bayesian approach easily incorporates the variability of the noise factors using the response modeling approach (Welch, Yu, Kang and Sacks 1990 and Shoemaker, Tsui and Wu 1991). This approach and its techniques are used to analyze two data sets, one which arises from a robust design experiment.

Key Words: Binary Data, Generalized Linear Model, Gibbs Sampler, Robust Design.

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 (three to six) is common. Also, a binary response is an important special case.

In industrial situations, such data may be cheap surrogates for continuous measurements. This occurs when the main priority is to run the experiment quickly, and if necessary a more complete measurement can be considered later. For example, one 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; instead an expert could 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 article have several features specific to industrial experiments. For example, the predictor variables (or factors) are typically set at a small number of levels, and are arranged in an orthogonal array. Also, the number of main effects and interactions is typically close to the number of observations made. This means that the models considered will be relatively simple, and identification of important effects and interactions will be a priority. Although only orthogonal arrays are considered in the examples, the methods developed will generally be applicable to arbitrary designs.

Interest focuses on modeling the way in which predictor variables influence the distribution of the ordered categorical response. Several such models are currently in use and are outlined in Section 2. These approaches have various difficulties such as inappropriate or inaccurate inference and infinite estimates of factor effects. A Bayesian model is proposed to circumvent these difficulties. Previously, Hamada and Wu (1995) showed the benefits of using a Bayesian approach for analyzing censored data because it avoids the problem of infinite estimates from standard methods.

The Bayesian model for ordered categorical data is described in Section 3 along with the Gibbs sampler, a computational technique used to fit the model. In Section 4, two industrial examples illustrate the analysis strategy using the Bayesian model. The strategy includes a model selection stage and an optimization stage in which factor levels are chosen. The Bayesian nature of the model makes it easy to model uncertainty in the parameters. Moreover, it can handle the variability of noise factors in Taguchi-style robust design experiments. In Section 5, a summary and conclusions are given.

2 Existing Analysis Techniques

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. In this section, these techniques and some of the problems they have are considered.

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 so-called *scored ANOVA* approach is simple, 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. As Nair (1986) mentioned, the main attraction of this technique is its simplicity, which suggests that other approaches could be more accurate.

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 articles 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 McCullagh's (1980) regression models for ordered categorical data. These Generalized Linear Models (GLMs) are attractive since software is readily available for fitting them. The method can be summarized as follows: instead of creating a pseudo measurement to be analyzed, the probability that an observation Y_i from a sample of size n will fall in category j is modeled as a function of the predictors. McCullagh (1980) suggests a family of models of the form:

$$\text{link}(\Pr(Y_i \leq j)) = \gamma_j - X_i'\boldsymbol{\beta} \qquad \text{for} \qquad j = 1, ..., J-1 \text{ and } i = 1, \ldots, n, \quad (1)$$

where categories are labeled $1, \ldots, J$, "link" is a (known) monotone increasing function mapping the interval (0,1) onto the real line $(-\infty, \infty)$, γ_j is a "cutpoint", X_i is a vector of covariates corresponding to 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 categories.

Once a link function is specified, estimation of the parameters $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ is carried out via maximum likelihood (ML). The likelihood is formed from (1), and its exact form is given in McCullagh (1980). One problem with maximum likelihood is that estimates for certain coefficients or cutpoints may be infinite. For example, consider 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 likely to occur, especially when the number of effects in the model is near the number of runs. Tse (1986) gives conditions under which ML estimates (MLEs) are infinite. In practice, these problems will manifest themselves as estimates which fail to converge.

Despite non-converging 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). However, in situations where the model is to be used for prediction or when comparisons between effects are necessary, infinite estimates are a problem. For example, if $\hat{\beta}' = (\hat{\beta}_1, \hat{\beta}_2) = (\infty, -\infty)$ then (1) is undefined at $X' = (1, 1)$.

One solution would be to assume some sort of prior knowledge of the coefficients, and use Bayesian techniques to fit a model to the data. Even when knowledge about the coefficients is minimal or nonexistent, this approach has justification; one suspects that the parameters are only large and that there is not sufficient data to distinguish between large and infinite values. On the probability scale such assumptions mean probabilities of exactly zero or one in (1) are unlikely, but values close to zero and one are plausible.

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 samples, where the asymptotic normality of ML theory may be a poor approximation. Details of this method are provided next.

3 A Bayesian Approach

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 marginal posteriors of the parameters of interest. A brief summary of the Gibbs algorithm is given before specific results are derived.

3.1 The Gibbs Sampler

The Gibbs sampler, which was introduced by Geman and Geman (1984), is a technique for the calculation of marginal distributions of random variables, given a set of conditional distributions. In the Bayesian context, it is used to obtain marginal posteriors of the parameters after conditioning on the observed data $\mathbf{Y}$. This is seen as a simpler alternative to numerical integration techniques, which may require specialized knowledge and attention to the details of a specific problem. The applications of this technique to statistical problems have been recently outlined in a number of articles, such as Gelfand and Smith (1990) and Smith and Roberts (1993).

To illustrate the algorithm, consider three random variables U_1, U_2, and U_3. The notation $f(\cdot)$ will denote the density of the argument. Thus, $f(X)$ is the density of X, $f(Y)$ is the density of Y, but $f(X) \neq f(Y)$.

The goal is to determine the marginal distributions $f(U_1), f(U_2)$, and $f(U_3)$. It is assumed that the distributions $f(U_i|U_j, j \neq i)$ are available (i.e., they may be sampled from). Each of these conditional distributions are called *full conditional distributions* because they condition on the value of all but one variable. Starting with an arbitrary set of values $(U_1^{(0)}, U_2^{(0)}, U_3^{(0)})$, $U_1^{(1)} \sim f(U_1|U_2^{(0)}, U_3^{(0)})$, $U_2^{(1)} \sim f(U_2|U_1^{(1)}, U_3^{(0)})$ and $U_3^{(1)} \sim f(U_3|U_1^{(1)}, U_2^{(1)})$ are sampled. The algorithm cycles repeatedly in the same fashion, always conditioning on the most recent values of (U_1, U_2, U_3). Geman and Geman (1984) show that these random variables converge in distribution to a sample from the joint distribution. Consequently, any subset of the three variables can be viewed as a sample from the appropriate marginal distribution.

There are several different strategies for implementing the Gibbs sampler; all relate to the central issue of obtaining a representative sample of the posterior. One important factor is the number of iterations required to remove the effect of the starting values (i.e. the burn-in time). For the examples given here the burn-in time can be quite large (500 to 1000 cycles for the example in Section 4.1), so it is most efficient to use many values taken from a single long run of the Gibbs sampler. The length of the burn-in time is determined by comparing several chains using different starting values. A second issue in obtaining a representative sample is the correlations between successive draws from a single chain. If correlations between successive draws are large, there is less information in the sample than its size indicates. Time series tools such as the autocorrelation function can be useful in determining the correlation, and the number of runs that are likely to be necessary. If storage space is a concern and correlations are large, it may be efficient to take only every mth sample after burn-in. In both examples here, all values from a single chain after burn-in are used. Details and further discussion of these issues is given in Section 4.

3.2 The Gibbs Sampler for Ordinal Data

The model to be fit is a Bayesian version of the GLM described in the previous section, with a probit link, a normal prior for the coefficients $\boldsymbol{\beta}$, and an ordered normal prior for the cutpoints $\boldsymbol{\gamma}$. The derivation of conditional distributions for the Gibbs sampler is similar to Albert and Chib (1993), but in this paper a different parameterization and informative priors are used. The ability to vary the shape of the prior will allow the robustness of the analysis to be assessed.

The form of the full conditional distributions is simplified by the assumption of an underlying continuous variable associated with the observed 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. Thus, 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 $\text{link}^{-1}(Z_i - X_i'\boldsymbol{\beta})$, since from (1), $\Pr(Z_i < \gamma_j) = \Pr(Y_i \leq j) = \text{link}^{-1}(\gamma_j - X_i'\boldsymbol{\beta})$. For a probit link, which uses the standard normal cdf as the link, $Z_i \sim N(X_i'\boldsymbol{\beta}, 1)$. In what follows, vector notation will be used, with $\mathbf{Y}, \mathbf{Z}, \boldsymbol{\gamma}$ as vectors of lengths n, n and $J-1$ respectively, and $\mathbf{X}$ denoting the design matrix.

Conditional on the observed $\mathbf{Y}$, 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, consider 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. Its pdf can be written as:

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

where the function $\mathcal{I}$ is 1 when $\gamma_{\mathbf{Y}_i - 1} \leq Z_i < \gamma_{\mathbf{Y}_i}$ for $i = 1, ..., n$ and 0 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. Thus,

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

for arbitrary densities $f(\boldsymbol{\beta}), f(\boldsymbol{\gamma}), f(\mathbf{Z}|\boldsymbol{\beta})$. As already mentioned, the link determines $f(\mathbf{Z}|\boldsymbol{\beta})$. The priors used here will be $\boldsymbol{\beta} \sim N(0, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$ and $\boldsymbol{\gamma} \propto N(0, \mathbf{D})$ where $\mathbf{D}$ is diagonal, and $\boldsymbol{\gamma}$ is restricted so that $\gamma_1 < \gamma_2 < \ldots < \gamma_{J-1}$. Typically $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ is also diagonal, with $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \sigma_\beta^2 \mathbf{I}$ for equally scaled predictors. In either case, diagonal elements are chosen so that many values of $\Pr(Y_i \leq j)$ are plausible. Details are given in Section 4.2.

The practical reason for these choices is that all the conditional distributions have forms that facilitate sampling. Nevertheless, the normal priors allow considerable flexibility and will help illustrate the application of these techniques without diverting too much attention to computational issues. The experiments considered are primarily screening experiments, so a prior mean of $\boldsymbol{\beta} = \mathbf{0}$ seems plausible. 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 the probabilities of the ordered categories (because of the probit link and taking $\sigma_{\gamma_j} = 1, j = 1, \ldots J-1$), and an uninformative prior on the underlying or transformed scale (i.e., taking $\sigma_{\gamma_j} \to \infty, j = 1, \ldots J-1$). By using a parameterization which puts all the intercepts in the same vector rather than having one in $\boldsymbol{\beta}$, separate and independent priors on $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ seem reasonable.

With these priors, the conditional distributions can readily be expressed as follows:

$$
\begin{aligned}
f(\boldsymbol{\beta}|\boldsymbol{\gamma}, \mathbf{Z}, \mathbf{Y}) &\propto f(\mathbf{Z}|\boldsymbol{\beta}) f(\boldsymbol{\beta}) \\
&= N((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}, (\mathbf{X}'\mathbf{X})^{-1}) \times N(0, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \qquad (2)\\
&= N((\boldsymbol{\Sigma}_{\boldsymbol{\beta}}{}^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}, (\boldsymbol{\Sigma}_{\boldsymbol{\beta}}{}^{-1} + \mathbf{X}'\mathbf{X})^{-1}), \\
f(\boldsymbol{\gamma}|\boldsymbol{\beta}, \mathbf{Z}, \mathbf{Y}) &\propto f(\boldsymbol{\gamma}) \mathcal{I}(\mathbf{Y}; \mathbf{Z}, \boldsymbol{\gamma}) \\
&= I(\gamma_1 < \ldots < \gamma_{J-1}) \prod_{j=1}^{J-1} N(0, \sigma^2_{\gamma_j}) \times \mathcal{I}(\mathbf{Y}; \mathbf{Z}, \boldsymbol{\gamma}) \qquad (3)\\
&= \prod_{j=1}^{J-1} N(0, \sigma^2_{\gamma_j}) \times I(\max_i\{Z_i | Y_i = j-1\} < \gamma_j < \min_i\{Z_i | Y_i = j\}), \\
f(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{Y}) &\propto f(\mathbf{Z}|\boldsymbol{\beta}) \mathcal{I}(\mathbf{Y}; \mathbf{Z}, \boldsymbol{\gamma}) \\
&= N(\mathbf{X}'\boldsymbol{\beta}, \mathbf{I}) \mathcal{I}(\mathbf{Y}; \mathbf{Z}, \boldsymbol{\gamma}) \qquad (4)\\
&= \prod_{i=1}^{n} N(X_i'\boldsymbol{\beta}, 1) I(\gamma_{Y_i - 1} < Z_i < \gamma_{Y_i}).
\end{aligned}
$$

The generic indicator function I is 1 whenever its argument is true and 0 otherwise. Note that in (2), $f(\mathbf{Z}|\boldsymbol{\beta})$ is viewed as a function of $\boldsymbol{\beta}$, while in (4), it is viewed as a function of $\mathbf{Z}$. In (2) the product of two normal pdfs is a normal pdf. The indicator functions in (3) 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 $\mathbf{Z}$ in (4) is similar with each element bounded by a pair of cutpoints. Notice that in both (3) and (4), 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.

Thus the Gibbs sampler will sample $\boldsymbol{\beta}$ all at once, and $\boldsymbol{\gamma}$ and $\mathbf{Z}$ one element at a time. The order of sampling is $\mathbf{Z}$, $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, and all the distributions are easy to sample. Note that if other priors were thought to be more 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, Inc. 1992).

4 Analysis Techniques

In this section, two industrial datasets 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 is of special interest in an industrial context, since usually one outcome is a target value. Methods for optimizing within the framework of Taguchi's robust design experiments will also be given.

4.1 A Foam Molding Experiment

The data, which was originally analyzed by Jinks (1987), arise from an experiment designed to reduce voids in a 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 1, and the data are given in Table 2. 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.

The initial 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 effects plus two cutpoints which define the three ordered categories. The model was fit with diffuse priors for $\boldsymbol{\gamma}$ ($\sigma_{\gamma_j} = \infty$) and a $N(\mathbf{0}, \mathbf{I})$ prior for $\boldsymbol{\beta}$ (i.e., $\sigma_{\beta} = 1$).

Before analyzing the posteriors, consider some details of the Gibbs algorithm for this application and an empirical method for assessing convergence. Several long chains from different starting values $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)$ were run and observed to see how long they took to reach the same distribution. In Figure 1, the behavior of four chains with starting values (marked with dots) $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) = \{(\mathbf{0}, (-5, -5))$

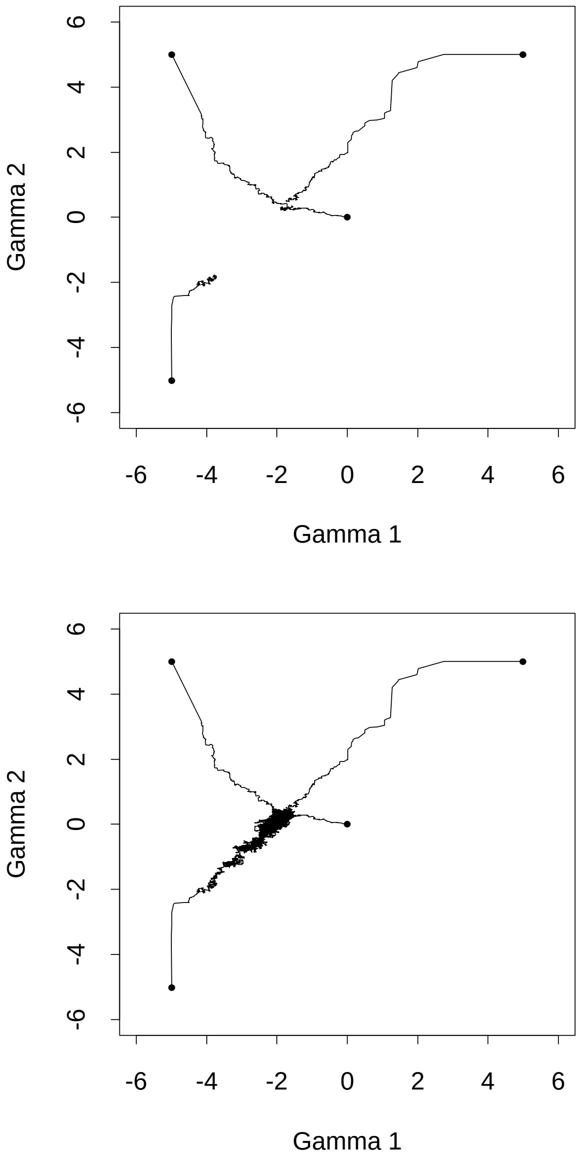

Figure 1: Convergence of γ, 250 and 5000 cycles.

	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 1: Foam molding experiment factors and levels.

						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

Table 2: Foam molding experiment design and frequencies for good (I), ok (II) and poor (III).

, $(\mathbf{0}, (-5, 5))$, $(\mathbf{0}, (5, 5))$, $(\mathbf{0}, (-0.1, 0.1))\}$ is illustrated. The first plot shows that the sequences had not converged by the 250th cycle. The second plot shows 5000 samples for the same chains and indicates convergence, which occurs somewhere between 500 and 1000 cycles. The convergent distribution appears small because of the large distance between different starting values. 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 (3) for $\boldsymbol{\gamma}$ and the large sample size (38, 156 and 126 responses in the three categories, respectively) explain the slow convergence, since γ_1 and γ_2 are restricted to lie between Z_i's corresponding to Y_i's of 1, 2 and 3. The random variation in the Z_i's leave very narrow regions for sampling $\boldsymbol{\gamma}$ at each cycle and thus convergence is slow (supported by examination of the autocorrelations, which die out slowly). Similar convergence problems are discussed in Cowles (1994) who proposes a way to accelerate the convergence. A single long run of the Gibbs sampler will be most efficient, and the informal diagnostics used above suggest the following strategy: use the last 10,000 samples from a single long chain of 11,000 cycles with starting values $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) = (\mathbf{0}, (-0.1, 0.1))$. Concern over the high correlations led to examination of a longer run (every 10th value from 100,000 cycles) of the chain, but since results were qualitatively the same, the smaller sample is used here.

The first analysis task is to identify the factors with the largest effects using the posteriors generated by the Bayesian analysis. The posteriors obtained via the Gibbs sampling algorithm are quite close to normal; hence a numeric summary of the marginal posteriors of the coefficients is given (in Table 3) rather than plots. Table 3 also summarizes the posteriors for γ_1 and γ_2.

The intention is that these and other summaries of the posteriors would be used to identify terms which could be dropped from the model. The posterior mean and standard deviation and quantiles all depend on the scaling of the predictor variables - in this case all variables are scaled to have levels ± 1 which means that their scales are comparable. As a supplement to the mean, standard deviation (std. dev.) and quantile information (not shown), the percentages of the posterior on each side of zero (zero splits) are also included. These summaries form the basis for the variable selection strategy presented here which iteratively fits models with decreasing numbers of terms.

In this case, the variables DH, EH, FH, GH, AI, BI, CI, FI and GI are dropped first and the model containing the remaining effects is refit. Several of the columns of Table 3 support this decision; their means are smaller than their corresponding standard deviations, and the zero splits indicate that the centers of the posteriors lie near zero. A conservative approach in dropping terms was taken. Because the process will be repeated, unimportant terms which escape removal at this stage will be removed at a subsequent iteration. The next iteration lead to the removal of the terms HI, BH, CH, DI, and D which left 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 so that they are not summarized here. It should be emphasized that this model selection approach is rather informal and that other approaches could also be used, such as automated Bayesian variable selection (see George and McCulloch 1993) or a predictive approach (Box 1980). While the approach relies on marginal posteriors, its stepwise nature means that the model space is more thoroughly explored than if a final model were decided upon using only the full model.

Table 3 also gives p-values for a scored ANOVA using the scores 1, 2 and 3. The results indicate differences between the ANOVA approach and the GLM. Specifically, ANOVA identifies D, DH, FH, and BI as significant, while the GLM does not. Although it is difficult to say which model is better on the basis of a single experiment, the GLM has several attractive features. This includes the intuitive concept of a latent variable and (in this case) a more parsimonious model.

Using the important effects identified above, optimization can now be con-

	Posterior				Scored ANOVA
	mean	std. dev.	Zero Split		p-value
			%<0	%>0	
A	-0.75	0.19	1.000	0.000	0.000
B	0.91	0.19	0.000	1.000	0.000
C	0.62	0.19	0.000	1.000	0.006
D	-0.24	0.19	0.866	0.134	0.019
E	-0.58	0.19	1.000	0.000	0.003
F	0.57	0.19	0.000	1.000	0.006
G	0.75	0.19	0.000	1.000	0.000
H	-0.45	0.17	1.000	0.000	0.000
I	0.37	0.16	0.000	1.000	0.000
HI	0.19	0.07	0.006	0.994	0.019
AH	0.43	0.17	0.000	1.000	0.001
BH	-0.20	0.17	0.896	0.104	0.830
CH	-0.24	0.17	0.945	0.055	0.668
DH	-0.03	0.17	0.590	0.410	0.003
EH	0.14	0.17	0.218	0.782	0.830
FH	-0.03	0.17	0.520	0.480	0.033
GH	-0.10	0.17	0.682	0.318	0.284
AI	-0.03	0.16	0.528	0.472	0.134
BI	-0.03	0.16	0.620	0.380	0.011
CI	0.11	0.16	0.270	0.730	0.830
DI	-0.18	0.16	0.892	0.108	0.830
EI	-0.36	0.16	0.999	0.001	0.003
FI	0.12	0.16	0.230	0.770	0.521
GI	0.09	0.16	0.325	0.675	0.199
γ_1	-2.08	0.21	1.000	0.000	
γ_2	-0.03	0.21	0.632	0.368	

Table 3: Foam experiment marginal posterior summaries and p-values for scored ANOVA.

sidered where the goal is to find settings which yield a high probability content in certain outcome categories. The vector of the probabilities of the ordered categories or outcome proportions $\mathbf{p}$ can also be thought of as a parameter with a posterior distribution since it is a function of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and X. To emphasize the distinction between the parameter $\mathbf{p}$ and its posterior, the terminology *proportion* $\mathbf{p}$ of outcomes and the posterior distribution of this proportion will be used. The posterior of $\mathbf{p}$ also depends on the predictors X, and these will be adjusted to find a "good" posterior for the proportions. There are many possible definitions of the term good, but the two goals of Taguchi come to mind, namely an on-target location and small variability. For ordered categorical data, this means that the multivariate distribution of the proportions has most mass in the category of interest with as little variation as possible.

The robust design aspect of this experiment adds a further challenge in the optimization process. In a robust design experiment, the predictors are divided into two groups: those which are controllable under regular production (control factors) and those which are expensive or impossible to control under production (noise factors). In the special off-line experiment that is carried out, both the control and noise factors are set at fixed levels, however. First, the response is modeled as a function of both the control and noise factors. Then the noise factors are assumed to follow a specific distribution. This allows control settings that desensitize the response to the variation in the noise factors to be chosen (Welch, Yu, Kang and Sacks 1990 and Shoemaker, Tsui and Wu 1991). Note that for the Bayesian model, the distribution of $\mathbf{p}$ now has two different components - the posterior for $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and the distribution of the noise factors. The calculation of the distribution of proportions is simple, and the additional distributions introduced for the noise factors are easily integrated into this calculation. First, values of the control factors are set at fixed levels. For each $(\boldsymbol{\beta}_l, \boldsymbol{\gamma}_l)$ in the posterior sample indexed by $l = 1, \ldots, N$, the noise factors are drawn from their distribution, and $\mathbf{p}$ is calculated using (1). Repeating this for the entire sample yields a sample from the distribution of proportions for a given setting of the control factors.

In the foam experiment, the variables H and I are the noise factors, so that settings of control factors A, B, C, E, F, G which maximize Pr(Good) need to be found. For illustrative purposes $(H, I) \sim N(\mathbf{0}, \mathbf{I})$ is assumed, although in practice this assumption would be based on expert knowledge. At each $(\boldsymbol{\beta}_l, \boldsymbol{\gamma}_l)$, H_l and I_l are drawn and Pr(Good) is calculated. The optimization is straightforward by considering all combinations of ± 1 levels for the six control factors. Using $E(\mathbf{p})$, promising settings can be identified, where the expectation is taken over $\boldsymbol{\beta}, \boldsymbol{\gamma}, H$ and I. The expected proportions for the $2^6 = 64$ different control settings are plotted in Figure 2. Only two components of $E(\mathbf{p})$ are plotted, since the three probabilities must sum to one. Thus, settings which have the most mass on Good and most of the remaining mass on OK are desirable – 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 –

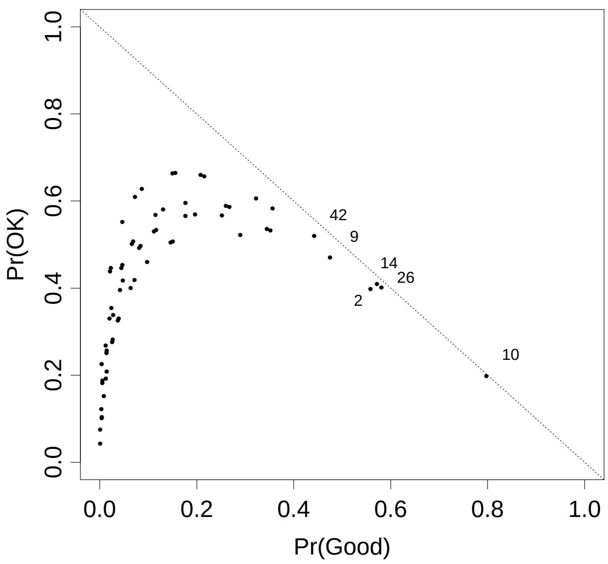

Figure 2: Expected proportions of good and ok parts at different settings of control factors A, B, C, E, F, G for the foam experiment.

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 of the mass is on the Good and OK categories so that $\sqrt{\text{Var}(\Pr(\text{Good}))} = 0.0873$ provides a good summary of variation. Because one setting is a clear winner, discussion of other interpretive tools for optimization will be postponed until the next example.

Before presenting the next example, an advantage of the Bayesian approach in optimization of robust design models needs to be emphasized. The use of a Bayesian model for optimization implies that uncertainty in $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ is accounted for in optimizing over the uncontrollable noise factors. In the robust design literature thus far, uncertainty has been ignored; rather, parameter estimates for the selected model are assumed to be "the truth". By including uncertainty in the optimization process, the Bayesian approach obtains more realistic estimates of one's knowledge of process performance.

4.2 An Injection Molding Experiment

This experiment, which is analyzed by Steinberg and Bursztyn (1993), 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 "higher-the-better" experiment previously discussed. The response variable is the "amount of material", ranging from 1 (too little) to 7 (too much), with a target value of 4. No values of 7 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 in a $2^{17-12} \times 2^2$ design. For illustrative purposes, consider the analysis of a quarter fraction of the original experiment with no noise effects (i.e., those data from mold #2 and water temperature of $8°C$). This subset of the original experiment 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.

The Gibbs algorithm converged much more rapidly for this problem, due to the smaller number of observations, which implies more freedom in the conditional distributions for $\boldsymbol{\gamma}$ (see the foam experiment for a detailed discussion). Autocorrelations died out much faster, suggesting that more than enough information would be available if the last 10000 values were used from a single chain of 10250 cycles.

The model fitting strategy is the same as in the previous example; a brief summary is given since the focus here will be on optimization. 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 A and AO terms were of borderline importance, but were included in the model.

One aspect of the model fitting that needs to be highlighted that was not

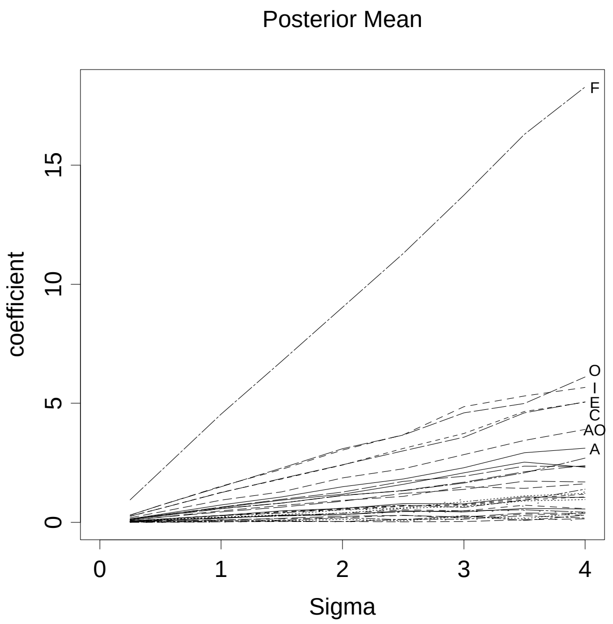

Figure 3: Injection molding experiment robustness of posteriors to prior variance.

mentioned in the previous example is robustness to the prior variance σ_β^2. Informative priors shrink the regression coefficient posteriors towards zero, thereby yielding finite parameter estimates. The degree of shrinkage depends on 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 (1995) and may be studied via a new graphical display, in which summaries of all marginal posteriors are plotted as a function of prior standard deviation σ_β. For example in Figure 3, the posterior mean of each marginal is plotted for prior $\sigma_\beta \in (0.25, 4)$. The magnitude of the summary changes over σ_β, but the rankings of factor effects do not. This robustness to prior information increases our confidence in the model selection process, which was carried out at a single value of σ_β $(= 1)$. This type of plot offers many possibilities such as studying the effect of changing the prior variance for one set of parameters with the prior variance for the remaining parameters held constant.

The rationale for using an upper limit of $\sigma_\beta = 4$ 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 $\sigma_\beta = 4$ is an adequately large upper bound.

The optimization stage of this example is interesting since challenges arise for the on-target problem that were not present in the foam molding experiment. 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 highest category. 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 other factors. For example, if one effect is positive, and the other negative, they could be used to cancel each other out to get on target. There are two ways to do this (if only factor settings of only ± 1 are considered) 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 can be chosen. Other secondary criteria can be used to distinguish between these settings.

These concepts are more clearly illustrated in the optimization stage for this experiment. Consider optimizing over all combinations of high and low levels (± 1) for all the factors or $2^6 = 64$ settings. All optimization is with respect to the posterior distributions for the proportions of parts in each response category. Thus, factor settings that produce on-target posteriors, namely ones with a high proportion in response group four are sought. That is, while there may be other favorable properties, it is crucial to produce a high proportion of on-target parts. To identify promising settings, therefore, consider the expected posterior proportion of parts in response category four. The five most promising settings are displayed in Table 4, along with posterior expected proportions and their standard deviations. The first four setting appear quite similar, and the fifth

#	**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)

Table 4: Injection molding experiment posterior expectations and standard deviations of (six response categories) proportions for five optimal settings.

has fewer on-target parts. The structure of the four best settings suggests choosing $F = 1, I = -1, O = 1, C = -E$ with the choice of A being apparently unimportant. If the goal is to have as many parts on target as possible, then a

(C, E) combination could be chosen on a cost basis. Other priorities can 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 example, there is little distinction between the first four settings, and the fifth has slightly more variation.

- **Other outcome categories.** 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 could be examined either by numeric methods or plots (not shown here). ¿From Table 4, settings 43-46 have 2% to 4% mass outside categories 3-5 while setting 58 has only 1%. If parts outside this range must be scrapped (and at great cost), these four settings would be less preferable.

- **A specific criterion based on projections.** The six dimensional distribution of proportions can 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 2 in the foam molding 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, since the products of proportions and profits for each outcome are summed. Thus, 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 into a "larger-the better" problem, since the goal is to maximize profit. The profit distribution as shown in Figure 4 indicates that 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 settings. Note that this specific scoring was chosen only for purposes of illustration.

One 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 category. The distinction is that projections are used here 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 our approach, the scores are added after a model has been identified.

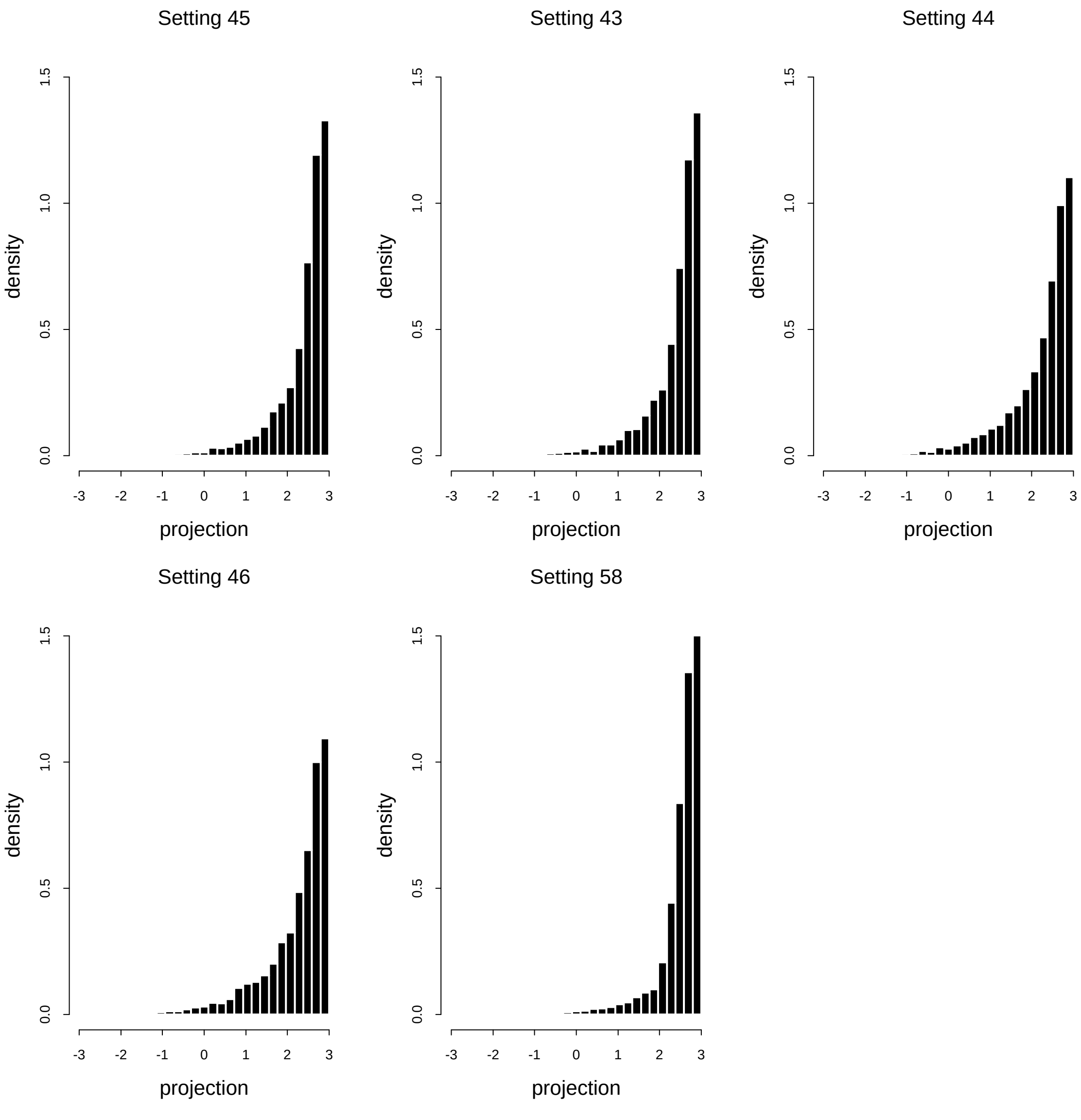

Figure 4: Injection molding experiment projected posterior proportions for projection
(-3,-3,-3,3,2,1) describing scrapped underfills and reworkable overfills.

Our philosophy is similar to the response modeling philosophy in robust design, which advocates fitting a model to the original data and not summaries of it. In robust design, the summaries are signal-to-noise ratios and the data are the original continuous responses; here the summaries are scores and the original data are category counts. Only after the model is fit are interpretive summaries then used. In response modeling, means and variances are used; here projections and other interpretive tools are employed.

5 Summary and Conclusions

For one reason or another, most of the current methods for analyzing ordered categorical data are problematic. The Bayesian approach outlined in this article is an attempt to overcome these difficulties. The combination of a Generalized Linear Model with Bayesian estimation techniques yields a model that appears appropriate for the data at hand. While the model itself has been considered previously, the area of application suggests new uses for the model, and advantages therein.

One of the most significant advantages is the ability of Bayesian models to account for uncertainty in the estimation of the parameters. By acknowledging the uncertainty rather than using point estimates, one hopefully gets a more honest choice of optimal factor settings. The uncertainty in the parameter estimates also fits in nicely in situations such as robust design experiments, where one assumes that additional variation in the process is induced by uncontrollable noise factors. The Bayesian approach provides an unified way of handling both sources of variability.

Other advantages of this approach include numerous graphical techniques for both model selection and process optimization, methods for checking the sensitivity of the results to the prior assumptions, and more meaningful conclusions, namely, an estimate of the proportion of parts in each category for a given factor level setting.

Of special interest is the on-target problem, since many different settings often produce similar results. In such situations, one has the option of either choosing the least expensive factor settings, or using other model-based criteria for distinguishing between competing settings. The number of possibilities is thereby reduced, which then needs to be confirmed by making a few additional runs.

While this approach deals with many difficulties previously encountered, the model considered is only one of many possible models that might be fit to such data. In some situations, it might be interesting to consider other Bayesian models for describing ordered categorical data, as well as adapting this model for a number of different experimental situations involving this type of data.

Acknowledgments

This research was supported by General Motors of Canada Limited, the Manufacturing Research Corporation of Ontario, and the Natural Sciences and Engineering Research Council of Canada. We thank C. F. J. Wu, G. Bennett and J. Lawless for insightful comments on an earlier version of the article. We also thank the associate editor and the referees whose helpful comments led to improvements in the current version.

References

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

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

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

Cowles, M. K. (1994), "Assessing and Accelerating Monte Carlo Markov Chain Convergence for Cumulative-Link Generalized Linear Models," Technical Report 94-001, Department of Preventative and Societal Medicine, University of Nebraska.

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.

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.

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," *Technometrics*, 32, 119–162.

Hamada, M. and Wu, C. F. J. (1995), "Analysis of Censored Data from Fractionated Experiments: a Bayesian Approach," *Journal of the American Statistical Association*, to appear.

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

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

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

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

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

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. 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.

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.

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.