Bayesian Analysis of Stochastically Ordered Distributions of Categorical Variables

Michael Evans Zvi Gilula Irwin Guttman Tim Swartz *

Abstract

This paper considers a finite set of discrete distributions all having the same finite support. The problem of interest is to assess the strength of evidence produced by sampled data for a hypothesis of a specified stochastic ordering among the underlying distributions and to estimate these distributions subject to the ordering. We present a Bayesian approach alternative to the use of the posterior probability of the hypothesis and the Bayes factor in favour of the hypothesis. Computational methods are developed for the implementation of Bayesian analyses. Examples are analyzed to illustrate the inferential and computational developments in the paper. The methodology employed for testing a hypothesis is seen to apply to a wide class of problems in Bayesian inference and has some distinct advantages.

*M. Evans is Professor, Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 1A1. Z. Gilula is Professor, Department of Statistics, Hebrew University, Jerusalem 91905, Israel. I. Guttman is Professor, Department of Statistics, State University of New York at Buffalo, Buffalo, New York 14214-3000, United States. T. Swartz is Associate Professor, Department of Mathematics and Statistics, Simon Fraser University, Burnaby, British Columbia, Canada V5A 1S6. M. Evans and T. Swartz were partially supported by grants from the Natural Sciences and Engineering Research Council of Canada.

Keywords : Bayes factors, Gibbs sampling, measures of concentration for posterior distributions, posterior probability, quadratic programming.

1. INTRODUCTION

In recent years the analysis of association amongst ordinal categorical data has received considerable attention. A list of pertinent references includes Agresti (1984), Feltz and Dykstra (1985), Goodman (1985, 1986, 1991), Gilula and Haberman (1986, 1988), Gilula and Ritov (1990) and Ritov and Gilula (1991, 1993). As argued by Gilula and Ritov (1990), ordinal constraints on categorical variables motivates the study of stochastic order between the distributions of such variables.

Let $\{F_i : 1 \leq i \leq m\}$ be a set of m distribution functions of categorical random variables all having the same support $\{1, \ldots, k\}$. Such a set of distributions is said to be stochastically ordered if $F_i(t) \geq F_{i+1}(t)$ for all t and for all i such that $1 \leq i \leq m-1$. If we let p_{ij} denote the probability of the response equaling j when the i-th distribution is true then this stochastic ordering imposes the following convex constraints on these probabilities; namely

$$\begin{aligned} p_{i1} &\geq p_{i+1,1} \\ p_{i1} + p_{i2} &\geq p_{i+1,1} + p_{i+1,2} \\ &\vdots \\ p_{i1} + \cdots + p_{ik} &\geq p_{i+1,1} + \cdots + p_{i+1,k} \end{aligned} \tag{1}$$

for $i = 1, \ldots, m-1$. Of course the p_{ij} also satisfy the constraints $p_{ij} \geq 0$ and $p_{i1} + \ldots + p_{ik} = 1$ for $i = 1, \ldots, m$. In this paper we are concerned with the situation where the p_{ij} are unknown and we have an independent sample from each of the F_i. We want to assess whether or not (1) holds and estimate the p_{ij} subject to (1).

Classical statistical inference for stochastic order between the distributions F_i is usually carried out in one of two ways. One approach fits a parametric model to the F_i where the order among the parameters determines the order between the distributions. Goodman's (1981) popular RC model is one of many examples of such models. Ritov and Gilula (1991) provide asymptotically efficient estimation techniques together with the corresponding testing procedures for the RC model when the parameters of the model are subject to relevant order constraints. The other approach is the model-free method. Here, efficient estimates (usually maximum likelihood) are obtained for the distributions F_i subject to a pre-specified stochastic order. Then appropriate goodness of fit tests are applied to test whether the data support such a pre-specified stochastic order. Relevant references for the model-free approach include Hanson et al. (1966), Grove (1980), Feltz and Dykstra (1985) and Kimeldorf et al. (1992) to name a few. Also see Robertson, Wright and Dykstra (1988) for an extensive discussion of order restricted inference methods.

In this paper we develop Bayesian methodology for assessing whether or not a particular stochastic order exists among the F_i. Further we compute Bayesian estimates of these distributions when a stochastic order has been determined to hold. The estimation context can arise in two distinct ways. First, the results of the testing might indicate that such a restriction is appropriate and second, we might be in a context where we are willing to assume (1) without testing. While this assumption may be somewhat unrealistic in general, there are examples, as pointed out in Feltz and Dykstra (1985), where this makes sense. For both situations the likelihood, as a function of the p_{ij}, takes the form of a product of m independent multinomials with a possibly restricted domain, depending on whether or not we assume (1) holds. We feel that the Bayesian approach has some advantages over the classical alternatives. For example, computing MLE's in this context can be a very difficult problem, see Feltz and Dykstra (1985). The Bayesian approach that we adopt avoids difficult optimization steps. Also Bayesian methodology does not require asymptotics for its justification. In the case of flat priors the Bayesian methods we develop are based on the form of the entire likelihood rather than just isolated characteristics such as a mode and curvature at the mode.

In Section 2 we consider the testing problem. Each $\mathbf{p}_{(i)} = (p_{i1}, \ldots, p_{ik})'$ is free to vary on the $(k-1)$-dimensional simplex, and we assume that the prior for each $\mathbf{p}_{(i)}$ follows a Dirichlet distribution and that these are independent. This specifies the prior distribution for the full parameter $\mathbf{p} = (\mathbf{p}'_{(1)}, \ldots, \mathbf{p}'_{(m)})$. The goal then is to assess whether or not (1) holds. Our approach generalizes that taken in Evans, Gilula and Guttman (1993) for the analysis of Goodman's RC model and can be seen as a natural generalization of the use of the posterior probability and the Bayes factor for assessing the evidence in favour of a hypothesis. This methodology can be applied to a much wider class of Bayesian hypothesis testing problems. In particular, the methodology handles the troublesome case where the null hypothesis is a lower dimensional subset of the parameter space. This results in prior and posterior probabilities equal to 0. A common recommendation in such a context is to modify the prior so that the null hypothesis has positive prior probability. Our approach avoids the need to do this.

In Section 3 we consider the estimation of $\mathbf{p}$. We consider two contexts for this. In the first situation we do not assume that (1) holds and present several estimators. In the second context we suppose that we are willing to assume the existence of the stochastic ordering given by (1) and then estimate $\mathbf{p}$. For this situation we take the prior to be as above but conditioned so that (1) holds.

In section 4 we present some examples. In section 5 we make some concluding remarks.

2. TESTING FOR STOCHASTIC ORDER

Suppose that we have m independent samples, one from each of the populations, with sample sizes $N_1, \ldots, N_m$ respectively. Let the number of reponses falling in the $j-$th cell for the $i-$th population be denoted by n_{ij}. Thus the data is $\mathbf{n} = (n_{11}, \ldots, n_{mk})'$. We then place a Dirichlet$(a_{i1}, \ldots, a_{ik})$ prior on the cell probabilities for each of the $i = 1, \ldots, m$ populations and assume independence. The posterior distribution of $\mathbf{p}$ is then given by

$$\mathbf{p}_{(i)} \sim \text{Dirichlet}(n_{i1} + a_{i1}, \ldots, n_{ik} + a_{ik}) \tag{2}$$

and these are independent for $i = 1, \ldots, m$. Therefore it is easy to directly sample from the posterior and obtain Monte Carlo estimates of various quantities. For example, if we generate $\mathbf{p}_1, \ldots, \mathbf{p}_N$ from the joint posterior we then estimate the posterior mean of p_{ij} by $\frac{1}{N}(p_{ij1} + \ldots + p_{ijN})$. See, for example, Devroye (1986) for a discussion of algorithms for generating from Dirichlet distributions.

Our first concern is to decide whether or not the data support the hypothesis that the constraints on the p_{ij} given by (1) hold. We denote the subset of $\mathbf{p}$ satisfying (1) by H_0 hereafter. One approach to this problem is to compute the posterior probability of H_0 and reject if this probability is very low. Alternatively the Bayes factor, given by the ratio of the posterior odds to the prior odds in favour of H_0, is computed with large values supporting H_0.

There is a significant and well-known problem, however, with these approaches in general. This is easily seen by considering a context, as in Evans, Gilula and Guttman (1993), where the posterior probability of H_0 is always zero irrespective of what data is obtained. In this case H_0 will always be rejected when the posterior probability is used and the Bayes factor is not defined. This situation typically arises when H_0 is a lower dimensional subset of the parameter space and so the prior probability of this set is 0 even though the prior density may indicate a relatively high degree of belief for values in this set. One way of dealing with this difficulty is to modify the prior so that it assigns a positive probability to H_0, as this ensures a nonzero posterior probability for H_0. While in many cases this produces an acceptable result, it is somewhat ad hoc as we may be required to change a perfectly reasonable prior for the full parameter. Accordingly it would be useful to have available a method for assessing the evidence in favour of H_0 that can handle hypotheses which are assigned zero prior probability without requiring modification of the prior.

We follow an approach taken in Evans, Gilula and Guttman (1993). As we will see this is a natural generalization of the use of the posterior probability and the Bayes factor described above but handles the situation where the hypothesis has 0 prior probability without requiring modification of the prior. This approach is based on assessing the degree to which the posterior distribution of $\mathbf{p}$ has increased its concentration about H_0 when compared to the concentration of the prior distribution of $\mathbf{p}$ about H_0. Clearly this concentration is only partially reflected, and sometimes not at all, in the prior and posterior probabilities of H_0. Accordingly we require the specification of a measure $\tau^2 = \tau^2(\mathbf{p}, H_0)$ of the distance of $\mathbf{p}$ from H_0. We then examine the prior and posterior distributions of τ^2. If the posterior distribution of τ^2 is much more concentrated near 0 than the prior distribution of τ^2 then we have evidence indicating that the hypothesis holds and not otherwise.

There are many possible choices for τ^2. A natural choice is based on least-squares; namely

$$\tau^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} (p_{ij} - x_{ij})^2 \tag{3}$$

where the p_{ij} are distributed according to (2), and for fixed p_{ij}, the $x_{ij} = x_{ij}(\mathbf{p})$ minimize (3) subject to (1). The minimization of (3) for $\mathbf{x}$ satisfying (1), is a quadratic programming problem and can be solved exactly using, for example, the IMSL subroutine QPROG. To derive the posterior distribution of τ^2 we use simulation; i.e. we generate $\mathbf{p}$ from (2), solve (3) for $\mathbf{x}$ and repeat this process many times. The prior distribution of τ^2 is computed in the same way. As another reasonable choice for τ^2 we could use Kullback-Liebler distance; i.e. for a given generated $\mathbf{p}$ calculate the $\mathbf{x}$ minimizing $\sum_{i=1}^{m} \sum_{j=1}^{k} p_{ij} \log(p_{ij}/x_{ij})$ subject to (1). Unless otherwise mentioned, however, our discussion will concern the least-squares distance measure.

We denote the posterior distribution function of τ^2 by $G(\cdot \mid \mathbf{n})$ and the prior distribution function by $G(\cdot \mid \mathbf{0})$. With this model and H_0, both the prior and posterior distributions of τ^2 are mixtures of discrete and continuous components as τ^2 takes the value 0 with positive prior and posterior probabilities. Thus in assessing the evidence for H_0 we are looking not only at these probabilities but also at how closely the continuous component of these distributions concentrates near 0. We assess the change in concentration from a priori to a posteriori by comparing $G(t \mid \mathbf{n})$ and $G(t \mid \mathbf{0})$ at values of t close to 0. The concentration of the prior distribution about H_0 represents our prior belief in H_0 holding approximately. The change in this concentration from a priori to a posteriori indicates how our belief in H_0 holding approximately has changed given the data. Therefore $G(\cdot \mid \mathbf{0})$ represents an appropriate standard by which to assess the concentration of the posterior distribution of τ^2. Note that $G(t \mid \mathbf{n})$ is the posterior probability of $H_t = \{\mathbf{p} : \tau^2(\mathbf{p}, H_0) \le t\}$; i.e. the posterior probability that the true value of $\mathbf{p}$ is within $t^{\frac{1}{2}}$ of some parameter value satisfying (1). Therefore our analysis generalizes the approach where one simply computes the posterior probability of the hypothesis H_0 when assessing the evidence in its favour. If $G(t \mid \mathbf{n})$ is high for small t then the model and data are assigning a high degree of posterior belief in H_0 holding approximately. If $G(t \mid \mathbf{n})$ is high relative to $G(t \mid \mathbf{0})$ for small t then the data has lead to a substantial increase in the degree of belief, from a priori to a posteriori, for H_0 holding approximately.

We can also generalize the Bayes factor approach to such problems by computing the Bayes factor of H_t as a function of t; namely

$$\mathrm{BF}(t) = \frac{G(t \mid \mathbf{n})/(1 - G(t \mid \mathbf{n}))}{G(t \mid \mathbf{0})/(1 - G(t \mid \mathbf{0}))}. \tag{4}$$

The value BF(0) is the usual Bayes factor in favour of H_0 although for problems where the numerator and denominator of (4) are both 0 this is not defined. We can, however, look at values of BF(t) for t close to 0 in all cases. As such this gives a unified treatment for testing hypotheses in a Bayesian context using Bayes factors and a fixed, given prior; i.e. hypotheses that have prior probability equal to 0 do not require special treatment. The interpretation of Bayes factors is somewhat less clear than posterior probabilities, however, and for that reason we prefer to assess the evidence in favour of a hypothesis by comparing $G(\cdot \mid \mathbf{n})$ to $G(\cdot \mid \mathbf{0})$. Plotting these functions seems like the most informative way of carrying out this comparison.

3. ESTIMATING THE STOCHASTICALLY ORDERED DISTRIBUTIONS

If, based on the preceding analysis, we decide that H_0 approximately holds we might then want to compute estimates of $\mathbf{p}$ subject to (1). When H_0 has positive posterior probability, as it does here, it makes some sense then to use the conditional posterior expectation of $\mathbf{p}$ given

H_0 as the estimate. To do this via simulation we need to be able to generate from a product of independent Dirichlets specified by (2), conditioned to H_0. Similarly if we assume (1) holds and take the prior to be a product of independent Dirichlets conditioned to H_0 then we are lead again to the same simulation problem. This problem is easily solved via the Gibbs sampling algorithm presented below but first we note some unsatisfactory features of this approach in general.

We note that conditioning on H_0 is equivalent to asserting the truth of H_0 while in general we may only be willing to conclude that H_0 holds approximately. Accordingly it would seem reasonable that the posterior probability mass not belonging to H_0 also affect our estimate. As another difficulty suppose that H_0 has prior probability 0 and thus has posterior probability 0 as well. In such a case the conditional posterior of $\mathbf{p}$ given H_0 is not defined. To do so requires the specification of a function defined on the parameter space that has H_0 as a preimage set and this can be done in numerous ways when there is not a natural choice imposed by the problem.

Several alternative estimators can be considered that avoid these problems. A natural estimate is to choose $\hat{\mathbf{c}} \in H_0$ that minimizes $\mathrm{E}[||\mathbf{p} - \mathbf{c}||^2]$ when $\mathbf{p}$ follows the posterior distribution given by (2). A simple calculation shows that $\hat{\mathbf{c}}$ minimizes $||\mathrm{E}[\mathbf{p}] - \mathbf{c}||^2$ and thus can be computed via quadratic programming. Recall that $\mathrm{E}[p_{ij}] = (n_{ij} + a_{ij}) / \sum_{k=1}^{m} (n_{ik} + a_{ik})$. A problem with this estimate, however, is that it will always lie on the boundary of H_0 whenever $\mathrm{E}[\mathbf{p}]$ is not an element of H_0. An alternative estimate that avoids all of these difficulties is given by $\hat{\mathbf{p}} = \mathrm{E}[\mathbf{x}(\mathbf{p})]$. Notice that $\hat{\mathbf{p}}$ equals the conditional expectation of $\mathbf{p}$ given H_0 when the posterior distribution is supported only on this set. Of course $\hat{\mathbf{p}}$ does not require the existence of the conditional distribution given H_0 for it to be defined. Further when H_0 is convex, as it is here, we have $\hat{\mathbf{p}} \in H_0$. When H_0 is not convex then some alternative characteristic of the posterior distribution of $\mathbf{x}$ must be chosen; e.g. the mode. When H_0 is a linear space then $\hat{\mathbf{p}} = \hat{\mathbf{c}}$. Numerically $\hat{\mathbf{p}}$ is computed by repeatedly generating $\mathbf{p}$, solving (3) for $\mathbf{x}$ and averaging these values. As a measure of the accuracy of this estimate it makes sense to look at $G(t \mid \mathbf{n})$. When this distribution is concentrated about 0 then we have evidence that $\hat{\mathbf{p}}$ is a good approximation to the true value of $\mathbf{p}$ and not otherwise. The quantity $\hat{\mathbf{p}}$ is our recommended estimator when we do not assume that (1) holds.

We consider now the situation where we estimate $\mathbf{p}$ assuming that H_0 holds. This could be due to actual physical constraints in the experiment or we have tested and decided that H_0 is reasonable. As discussed above we are required to sample from a product of independent Dirichlets conditioned to H_0. It does not seem possible to simulate exactly from this posterior distribution. The Gibbs sampler, however, which we show can be easily implemented in this context, permits the almost sure estimation of posterior distribution characteristics. If $\mathbf{p}_1, \ldots, \mathbf{p}_N$ constitutes N iterations of the Gibbs sampler then we estimate the posterior mean of p_{ij} by $\frac{1}{N}(p_{ij1} + \ldots + p_{ijN})$ where p_{ijt} is the t^{th} generated value of p_{ij}. We denote these estimates of the posterior means of the p_{ij} by p^*_{ij}. For a review of the use of the Gibbs sampler and related issues, see Smith and Roberts (1993) and Gelfand, Hills, Racine-Poon and Smith (1990) for an illustration of Bayesian applications in normal models.

The Gibbs sampler requires that we be able to sample from the conditional posterior distribution of each p_{ij} given H_0 and all the remaining coordinates of $\mathbf{p} = (p_{11}, \ldots, p_{mk})'$. We observe that the posterior density of $\mathbf{p}$ is proportional to

$$\prod_{i=1}^{m} \prod_{j=1}^{k} {p_{ij}}^{n_{ij}+a_{ij}-1} \tag{5}$$

where $p_{ij} \geq 0, p_{i1} + \ldots + p_{ik} = 1$ for $i = 1, \ldots, m$ and the constraints given by (1) are also satisfied.

Before writing down the conditional density of p_{ij} given H_0 and the remaining coordinates of $\mathbf{p}$ we need to introduce some notation. First we define the distribution p_{0j} for $j = 1, \ldots, k$ by requiring $p_{01} = 1$ and also define the distribution $p_{m+1,j}$ for $j = 1, \ldots, k$ by requiring $p_{m+1,k} = 1$. Note that these distributions will obey the stochastic ordering by the first index as expressed in (1). Now let $S_{j,l}^{(i)} = p_{ij} + p_{i,j+1} + \cdots + p_{il}$ for $i = 0, \ldots, m+1$ and $1 \leq j \leq l \leq k$ and be equal to 0 otherwise. From the inequalities in (1) and $p_{ij} \geq 0$, $p_{i1} + \ldots + p_{ik} = 1$ we obtain straightforwardly that this density is concentrated on $[l_{ij}, u_{ij}]$ where $l_{ij} = \max\{0, S_{1,j}^{(i+1)} - S_{1,j-1}^{(i)}, S_{1,j+1}^{(i+1)} - S_{1,j-1}^{(i)} - S_{j+1,j+1}^{(i)}, \ldots, S_{1,k-1}^{(i+1)} - S_{1,j-1}^{(i)} - S_{j+1,k-1}^{(i)}\}$ and $u_{ij} = \min\{S_{1,j}^{(i-1)} - S_{1,j-1}^{(i)}, S_{1,j+1}^{(i-1)} - S_{1,j-1}^{(i)} - S_{j+1,j+1}^{(i)}, \ldots, S_{1,k-1}^{(i-1)} - S_{1,j-1}^{(i)} - S_{j+1,k-1}^{(i)}\}$. We have from (5) that the conditional posterior density of p_{ij} given H_0 and all of the other cell probabilities is proportional to ${p_{ij}}^{n_{ij}+a_{ij}-1}(1 - S_{1,j-1}^{(i)} - S_{j+1,k-1}^{(i)} - p_{ij})^{n_{ik}+a_{ik}-1}$ for $i = 1, \ldots m$, $j = 1, \ldots k-1$ and $p_{ik} = 1 - \sum_{j=1}^{k-1} p_{ij}$. Setting $q_{ij} = S_{1,j-1}^{(i)} + S_{j+1,k-1}^{(i)}$ it follows immediately that the conditional posterior distribution function F_{ij} of p_{ij} given H_0 and the other coordinates of $\mathbf{p}$ is given by

$$\frac{B(p_{ij}/(1-q_{ij}), n_{ij}+a_{ij}-1, n_{ik}+a_{ij}-1) - B(l_{ij}/(1-q_{ij}), n_{ij}+a_{ij}-1, n_{ik}+a_{ij}-1)}{B(u_{ij}/(1-q_{ij}), n_{ij}+a_{ij}-1, n_{ik}+a_{ij}-1) - B(l_{ij}/(1-q_{ij}), n_{ij}+a_{ij}-1, n_{ik}+aij-1)} \tag{6}$$

where $B(y, a, b)$ is the probability that a Beta(a, b) variate is less than y. Therefore, given a variate $u \sim$ Uniform$(0, 1)$, and a subroutine for calculating the distribution function of Beta random variables, we could use bisection to generate from (6) via $F_{ij}(p_{ij}) = u$. Alternatively, we can generate from (6) using the adaptive rejection sampling algorithm for log-concave densities, as developed in Gilks and Wild (1992), as the doubly truncated Beta density is log-concave. This algorithm is by far the more efficient method for generating from these distributions.

For both Monte Carlo methods described in this paper convergence was assessed by comparing estimates of posterior distribution characteristics for increasing Monte Carlo sample size. While this is not foolproof there is currently no method of assessing convergence which is much more reliable than this. Of course this is a characteristic of virtually all iterative numerical procedures. As we will see in our examples the techniques described here produced meaningful accuracies within reasonable computation times.

4. EXAMPLES

Example 1. We consider a dataset taken from Srole, Langner, Michael, Opler and Rennie (1962) which investigates the relationship between an individual's mental health status ($m = 4$) and the socioeconomic status of the parents ($k = 6$). The data is displayed in Table 1. This example was also examined in Goodman (1985), Gilula (1986) and Evans, Gilula and Guttman (1993). For the prior we use a uniform prior; i.e. $\mathbf{p}_{(i)} \sim \text{Dirichlet}(1, \ldots, 1)$ and these are independent for $i = 1, \ldots, m$.

Sampling directly from the appropriate distribution in (2) we compute $G(\cdot \mid \mathbf{n})$, $G(\cdot \mid \mathbf{0})$, the estimate $\hat{\mathbf{p}}$ and the posterior standard deviations of the x_{ij}. A Monte Carlo sample of size 10,000 is used and provides 3 decimal places of accuracy. These computations require approximately 3 minutes of computing time on a Sun Sparcstation.

In Figure 1 we plot $G(\cdot \mid \mathbf{n})$ and in Figure 2 we plot $G(\cdot \mid \mathbf{0})$ together with $G(\cdot \mid \mathbf{n})$. In Table 2 we give some quantiles of these distribution functions. Further the posterior mean of τ^2 is .0004 while the prior mean is .1520. It seems clear from the plots and the table that the posterior distribution of τ^2 is highly concentrated near 0 relative to the prior distribution of τ^2 and we conclude that the data strongly support the stochastic ordering given by (1). Notice that in Figure 2 the posterior distribution function of τ^2 is virtually a vertical and then horizontal line indicating that relatively all the posterior probability is distributed near 0 when compared to the prior distribution.

We also have $G(0 \mid \mathbf{n}) = .1596$ and $G(0 \mid \mathbf{0}) = .0002$. Thus there is a large increase in the probability allocated to the submodel of stochastic ordering when we take the data into account. Still, the posterior probability by itself is not very convincing evidence in favour of the hypothesis. The Bayes factor is $\text{BF}(0) = 949$ and this seemingly provides strong support for the hypothesis. Also values of $\text{BF}(t)$ for t close to 0 give strong support to the hypothesis; e.g. $\text{BF}(.005) =$

592501. We note, however, that the comparison of $G(\cdot \mid \mathbf{n})$ and $G(\cdot \mid \mathbf{0})$ is overwhelmingly convincing that the posterior distribution of $\mathbf{p}$ provides support for the hypothesis given by (1). As such there seems to be little need to make use of these Bayes factors here.

In Table 3 we give the values of $\hat{p}_{ij}$; i.e. the posterior means of the x_{ij}. In Table 4 we give the posterior standard deviations of the x_{ij}. The standard errors of these estimates are all bounded by .0002. It is a simple matter to obtain the marginal posterior distribution of any x_{ij} just as we did for τ^2.

As discussed in Section 3 Gibbs sampling can be used to compute the conditional posterior means of the p_{ij} given that we are willing to assume that (1) holds; i.e. compute the value of $\mathbf{p}^*$. Table 5 gives the results from 50,000 iterations of the Gibbs sampling process and again it is felt that these are accurate to 3 decimal places. Of some note is the striking similarity between the estimates in Tables 3 and 5. This provides further support for our belief that a stochastically ordered model is correct.

Of some interest is the method used to generate from the distributions given by (6). To generate the 50,000 iterates using the bisection algorithm, where the equation $F_{ij}(p_{ij}) = u$ is solved to a tolerance of .000001, requires 7 hours and 51 minutes of computing time. This is the inversion method of generating random variables. However, the adaptive rejection algorithm method of Gilks and Wild (1992) reduces the computing time to only 24 minutes. This represents an improvement by a factor of about 20.

Example 2. Using the same data and model as in Example 1 we replace the least-squares measure of concentration by that based on the Kullback-Liebler distance measure. The results of the analysis are very similar. Once again comparing $G(\cdot \mid \mathbf{n})$ and $G(\cdot \mid \mathbf{0})$ leads overwhelmingly to the conclusion that H_0 holds. For example, the posterior mean of τ^2 is .0014 compared to the prior mean of .4404. Also the $\hat{p}_{ij}$ are almost identical with those recorded in Table 3.

Example 3. In this example we investigate the sensitivity of our testing approach via a simulation study. Table 6 contains the population parameters for $m = 3$ populations and $k = 5$ categories. The first 2 populations are stochastically ordered but the stochastic ordering fails in the penultimate cell for the last population. We carry out the test for stochastic order using simulated data from these populations for samples of sizes $N_1 = N_2 = N_3 = 50, 100, 500$. We use the least-squares distance measure and a uniform prior. In Table 7 we present the prior and posterior probabilities for H_0 and some prior and posterior quantiles for the distributions of τ^2 for the simulated data sets based on Monte Carlo samples of size 10,000. We see from this that our method is better at detecting lack of stochastic order as the sample size grows. Note that the posterior probability should decrease with N_i and the quantiles should increase. Of course, due to sampling variation, this will not strictly hold for every simulated data set.

Example 4. Throughout the paper and the previous examples we have always assumed that there is a specific stochastic ordering that we wish to check for. In many contexts, however, we may feel that some stochastic ordering exists but do not have a clear a priori idea of what it might be. In such a situation it makes sense to examine all possible stochastic orderings and determine which are the most plausible. Of course there are $m!$ different possible stochastic orderings so this technique is only feasible for relatively small m.

In Table 8 we present the data from a survey of car dealers measuring satisfaction levels with how the manufacturer handles spare parts. We restrict our analysis to Japanese cars grouped by maker where Other stands for Mitsubishi, Mazda and Isuzu. Satisfaction levels range from 5 (very satisfied) to 1 (very dissatisfied). Looking at all 24 possible stochastic orderings using our aproach the four most plausible stochastic orderings are recorded in Table 9. Clearly there is strong evidence for the ordering (Honda, Toyota, Other, Nissan) and thus Honda provides superior satisfaction.

5. CONCLUSIONS

We have examined the Bayesian analysis of stochastic ordering among a set of categorical variables. To assess the evidence in favour of a hypothesis, as summarized by a posterior distribution, we have introduced a measure of concentration of the posterior distribution about the hypothesis. Our approach is a natural generalization of the method where evidence is assessed by the use of the posterior probability of the hypothesis. Further, as shown for example in Evans, Gilula and Guttman (1993), our method handles the situation where the hypothesis always has posterior probability zero. The methodology clearly has applications in a wide variety of Bayesian hypothesis testing problems.

Of course there are different measures of concentration that can be chosen and we have used the intuitively reasonable measure of least-squares. Other than the need to implement a possibly troublesome minimization step, there is essentially no reason why other measures could not be chosen. For any given problem one could investigate the issue of sensitivity to detecting violations of the hypothesis to determine which is most useful in a given application. Presumably a particular measure of concentration will be more adept at detecting certain kinds of departures from the hypothesized submodel than others but we do not pursue this issue further here.

REFERENCES

Agresti, A. G. (1984), *Analysis of Ordinal Categorical Data*, J. Wiley, New York.

Devroye, L. (1986), *Non-Uniform Random Variate Generation*, Springer-Verlag, New York.

Evans, M., Gilula, Z., Guttman, I. (1993), "Computational issues in the Bayesian analysis of categorical data: log-linear and Goodman's RC model", *Statistica Sinica*, Vol. 3, No. 2, 391-406.

Feltz, C.J. and Dykstra, R.L. (1985), "Maximum likelihood estimation of the survival functions of N stochastically ordered random variables", *Journal of the American Statistical Association*, Vol. 80, No. 392, 1012-1019.

Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith, A. F. M. (1990), "Illustration of Bayesian inference in normal data models using Gibbs sampling", *Journal of the American Statististical Association*, 85, 972-985.

Gilks, W. R. and Wild, P. (1992), "Adaptive rejection sampling for Gibbs sampling", *Applied Statistics*, 41, 337-348.

Gilula, Z.(1986), "Grouping and association in contingency tables : an exploratory canonical correlation approach", *Journal of the American Statistical Association*, 81, 773-779.

Gilula, Z. and Haberman, S. J. (1986), "Canonical analysis of two-way contingency tables by maximum likelihood", *Journal of the American Statistical Association*, 81, 780-788.

Gilula, Z. and Haberman, S. J. (1988), "The analysis of multivariate contingency tables by restricted canonical and restricted association models", *Journal of the American Statistical Association*, 83, 760-771.

Gilula, Z. and Ritov, Y. (1990), "Inferential ordinal correspondence analysis: motivation, derivation and limitations", *International Statistical Review*, 58, 99-108.

Goodman, L. A. (1981), "Association models and canonical correlation in the analysis of cross-classification having ordered categories", *Journal of the American Statistical Association*, 76, 320-334.

Goodman, L. A. (1985), "The analysis of cross-classified data having ordered or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries", *Annals of Statistics*, 13, 1-9.

Goodman, L. A. (1986), "Some useful extensions of the usual correspondence analysis approach and the usual log-linear approach in the analysis of contingency tables", *International Statistcal Review*, 54, 254-309.

Goodman, L. A. (1991), "Measures, models and graphical displays in the analysis of cross-classified data (with comments and rejoinder)", *Journal of the American Statistical Association*, 86, 1085-1138.

Grove, D. M. (1980), "A test of independence against a class of ordered alternatives in a $2 \times c$ contingency table", *Journal of the American Statistical Association*, 75, 454-459.

Hanson, D. L., Brunk, H. D., Franck, W. E. and Hogg, R. V. (1966), "Maximum likelihood estimation of the distributions of two stochastically ordered random variables", *Journal of the American Statistical Association*, 61, 1067-1080.

Kimeldorf, G., Sampson, A. R. and Whitaker, L. R. (1992), "Min and max scorings for two-sample ordinal data", *Journal of the American Statistical Association*, 87, 241-247.

Ritov, Y. and Gilula, Z. (1991), "The order restricted RC model for ordered contingency tables: Estimation and testing for fit", *Annals of Statistics*, 19, 2090-2101.

Ritov, Y. and Gilula, Z. (1993), "Analysis of contingency tables by correspondence models subject to order constraints", To appear in *Journal of the American Statistical Association*

Robertson, T., Wright F.T. and Dykstra, R.L. (1988), *Order Restricted Statistical Inference*, John Wiley and Sons, New York.

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.

Srole, L., Langner, T.S., Michael, S.T., Opler, M.K. and Rennie, T.A.C. (1962), *Mental Health in the Metropolis: The Midtown Manhattan Study*, McGraw-Hill, New York.

Table 1: Cross-classification of subjects by mental health status and socioeconomic status of the parents in Example 1.

	Parents' Socioeconomic Status					
Mental Health Status	A	B	C	D	E	F
Well	64	57	57	72	36	21
Mild Symptoms	94	94	105	141	97	71
Moderate Symptoms	58	54	65	77	54	54
Impaired	46	40	60	94	78	71

Figure 1: Posterior distribution function of τ^2 in Example 1.

Figure 2: Posterior ... and prior — distribution functions of τ^2 in Example 1.

Table 2: Quantiles corresponding to posterior and prior probability distributions of τ^2 in Example 1.

p	Posterior	Prior
.05	0.0000	0.0270
.10	0.0000	0.0430
.25	0.0000	0.0795
.50	0.0002	0.1332
.75	0.0006	0.2049
.90	0.0013	0.2840
.95	0.0018	0.3383

Table 3: The estimates $\hat{p}_{ij}$ in Example 1.

	j=1	j=2	j=3	j=4	j=5	j=6
i=1	0.208	0.185	0.185	0.233	0.118	0.071
i=2	0.163	0.158	0.175	0.231	0.158	0.115
i=3	0.153	0.148	0.179	0.215	0.153	0.152
i=4	0.119	0.104	0.154	0.240	0.200	0.183

Table 4: Posterior standard deviations of the x_{ij} in Example 1.

	j=1	j=2	j=3	j=4	j=5	j=6
i=1	0.022	0.022	0.022	0.024	0.018	0.014
i=2	0.013	0.014	0.014	0.016	0.014	0.013
i=3	0.014	0.017	0.019	0.020	0.017	0.016
i=4	0.016	0.015	0.018	0.021	0.020	0.019

Table 5: The estimates p^*_{ij} in Example 1.

	j=1	j=2	j=3	j=4	j=5	j=6
i=1	0.210	0.185	0.185	0.233	0.118	0.069
i=2	0.166	0.159	0.175	0.229	0.157	0.115
i=3	0.147	0.145	0.178	0.219	0.157	0.154
i=4	0.116	0.103	0.154	0.240	0.200	0.187

Table 6: The population parameters p_{ij} in Example 3.

i	j=1	j=2	j=3	j=4	j=5
1	.20	.20	.20	.20	.20
2	.15	.12	.11	.10	.52
3	.14	.11	.11	.20	.44

Table 7: The prior and posterior probability of H_0 and some prior and posterior quantiles of the posterior distribution of τ^2 in Example 3.

	Probability of H_0	.01 quantile	.05 quantile
Prior	.0121	.0000	.0003
Posterior $N_i = 50$	.0620	.0000	.0000
Posterior $N_i = 100$	.0007	.0013	.0046
Posterior $N_i = 500$	.0000	.0054	.0080

Table 8: Data for Japanese Car Manufacturer Satisfaction Study where 5 is very satisfied and 1 is very dissatisfied in Example 4.

Satisfaction level	5	4	3	2	1
Honda	85	64	18	6	5
Toyota	67	44	27	12	4
Nissan	34	38	36	18	26
Other	82	83	74	28	24

Table 9: Prior and posterior quantities for the most plausible stochastic orderings where H = Honda, T = Toyota, N = Nissan and O = Other in Example 4.

Order	$G(0 \mid \mathbf{0})$	$G(0 \mid \mathbf{n})$	Prior mean of τ^2	Posterior mean of τ^2	BF(.0005)
HTON	.0004	.3125	.1840	.0004	3858.8
THON	.0004	.0020	.1840	.0070	23.3
HTNO	.0004	.0000	.1840	.0096	2.5
THNO	.0004	.0000	.1840	.0163	0.0