# Bayesian Analysis of Ordinal Survey Data using the Dirichlet Process to Account for Respondent Personality Traits

Saman Muthukumarana and Tim B. Swartz *

**Abstract**

This paper presents a Bayesian latent variable model used to analyze ordinal response survey data by taking into account the characteristics of respondents. The ordinal response data are viewed as multivariate responses arising from continuous latent variables with known cut-points. Each respondent is characterized by two parameters that have a Dirichlet process as their joint prior distribution. The proposed mechanism adjusts for classes of personalities. The model is applied to student survey data in course evaluations. Goodness-of-fit (gof) procedures are developed for assessing the validity of the model. The proposed gof procedures are simple, intuitive and do not seem to be a part of current Bayesian practice.

**Keywords** : Dirichlet process, Goodness-of-fit, latent variables, MCMC, WinBUGS.

*Saman Muthukumarana is Assistant Professor, Department of Statistics, University of Manitoba, Winnipeg Manitoba, Canada R3T2N2. Tim Swartz is Professor, Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby British Columbia, Canada V5A1S6. Both authors have been partially supported by research grants from the Natural Sciences and Engineering Research Council of Canada. The authors thank two anonymous reviewers whose comments led to an improvement in the manuscript.

# 1 INTRODUCTION

For the sake of convenience, many surveys consist of ordinal data, often collected on a five-point scale. For example, in a typical course evaluation survey, a student may express his view concerning an aspect of the course from a set of five alternatives: 1-poor, 2-satisfactory, 3-good, 4-very good, and 5-excellent. Sometimes five-point scales have alternative interpretations. For example, the symmetric Likert scale measures a respondent's level of agreement with a statement according to the correspondence: 1-strongly disagree, 2-disagree, 3-neither agree nor disagree, 4-agree, and 5-strongly agree. Student feedback on course evaluation surveys represents a modern approach for measuring quality. Nowadays, a growing number of websites use student feedback as their main performance indicator in teaching evaluations. As an example, http://www.ratemyprofessors.com/ rate over one million professors based on student feedback on a five-point ordinal scale. The scenario is similar in customer satisfaction surveys and social science surveys.

The simplest method of summarizing ordinal response data is to report the means corresponding to the ordinal scores for each survey question. At a slightly higher level of statistical sophistication, standard ANOVA methods may be applied to the ordinal scores by treating the data as continuous. However, the standard models for the analysis of ordinal data are logistic and loglinear models (Agresti 2010, McCullagh 1980 and Goodmann 1979). These models correctly take into account the true measurement scales for ordinal data and permit the use of statistical inference procedures for assessing population characteristics. An overview of the methodologies for ordered categorical data is given by Liu and Agresti (2005).

The approach in this paper is Bayesian and considers an aspect of ordinal survey data that is sometimes overlooked. It is widely recognized that respondents may have differing personalities. For example, consider a company which conducts a customer satisfaction survey where there is a respondent with a negative attitude. The respondent may complete the survey with a preponderance of responses in the 1-2 range. In this case, a response of 1 may not truly represent terrible performance on the part of the company. The response may reflect more on the disposition of the individual than on the performance of the company. As another example of an atypical personality, consider an individual who only provides extreme responses of 1's and 5's. It would be useful if statistical analyses could adjust for personalities. This is the motivation of the paper, and the tool which we use to account for personalities is the Dirichlet process, first introduced by Ferguson (1973). As a by-product of the proposed methodology, we attempt to identify areas (survey questions) where performance has been poor or exceptional. In addition, we attempt to identify questions that are highly correlated. Clearly, surveyors desire accurate responses and by identifying highly correlated questions, it allows surveyors to remove redundant questions from the survey which in turn reduces fatigue on the part of the respondents.

Our paper is not the first Bayesian paper to consider this problem. Alternative Bayesian approaches include Johnson (1996), Johnson (2003), Dolnicar and Grun (2007), Rossi, Gilula and Allenby (2001), Kottas, Mueller and Quintana (2005), Javaras and Ripley (2007) and Emons (2008). Johnson (2003) uses a hierachical ordinal regression model with heterogenious thresholds structure. Dolnicar and Grun (2007) use a ANOVA approach to assess the inter-cultural differences in responses. Rossi, Gilula and Allenby (2001) address nonidentifiability and parsimony by imposing various complex constraints on the unknown cut-points. Kottas, Mueller and Quintana (2005) propose a nonparametric Bayesian approach to model multivariate ordinal data recorded in contingency tables. One of the main features of this paper is that there is a mechanism to cluster subjects based on personalities. Most importantly, in our approach, clustering takes place as a part of the model and data determine the clustering structure. Often, clustering is done in a post hoc fashion, following some fitting procedure.

In addition to the methodological contribution provided in this paper, issues related to scaling are also considered. Not only does the approach attempt to remove idiosyncratic scaling, assumptions are made about the manner in which individuals transform latent continuous scores to discrete scores. There is a considerable literature on the psychology of survey response, the impact of survey question format, the effect of scales, etc. For a brief introduction to some of these topics, the reader is referred to Tourangeau et al. (2000), Fanning (2005) and Dawes (2008). For an introduction to the analysis of ordinal data in the applied fields of education and medicine, the reader is referred to Cohen, Manion and Morrison (2007), and Forrest and Andersen (1986) respectively.

In section 2, we provide a detailed development of the Bayesian latent variable model proposed in the paper. The model assumes that ordinal response data arise from continuous latent variables with known cut-points. Furthermore, each respondent is characterized by two parameters that have a Dirichlet process as their joint prior distribution. The mechanism adjusts for classes of personalities leading to standardized scores for respondents. Prior distributions are defined on the model parameters. We provide details about nonidentiability in our model and we overcome nonidentifiability issues by assigning suitable prior distributions. Computation is discussed in section 3. As the resulting posterior distribution is complex and high-dimensional, we approximate posterior summary statistics which describe key features in the model. In particular, posterior expectations are obtained via MCMC methods using WinBUGS software (Spiegelhalter, Thomas and Best 2003). In section 4, the model is applied to actual student survey data obtained in course evaluations. A comparison is made with an analysis based on the methodology of Rossi, Gilula and Allenby (2001). We then demonstrate the reliability of the approach via simulation. In section 5, goodness-of-fit procedures are developed for assessing the validity of the model. The proposed procedures are simple, intuitive and do not seem to be a part of current Bayesian practice. We conclude with a short discussion in section 6.

## 2 MODEL DEVELOPMENT

Consider a survey where the observed data are described by a matrix $X : (n \times m)$ whose entries $X_{ij}$ are the ordinal responses. The $n$ rows of $X$ correspond to the individuals who are surveyed and the $m$ columns refer to the survey questions. Without loss of generality, we assume that the responses are taken on a five-point scale.

We assume that the discrete response $X_{ij}$ of individual $i$ to survey question $j$ arises from an underlying continuous variable $Y_{ij}$. We consider a cut-point model which converts the latent variable $Y_{ij}$ to the observed $X_{ij}$ as follows:

$$\begin{aligned}
X_{ij} = 1 &\Longleftrightarrow \lambda_0 < Y_{ij} \leq \lambda_1 \\
X_{ij} = 2 &\Longleftrightarrow \lambda_1 < Y_{ij} \leq \lambda_2 \\
X_{ij} = 3 &\Longleftrightarrow \lambda_2 < Y_{ij} \leq \lambda_3 \\
X_{ij} = 4 &\Longleftrightarrow \lambda_3 < Y_{ij} \leq \lambda_4 \\
X_{ij} = 5 &\Longleftrightarrow \lambda_4 < Y_{ij} \leq \lambda_5
\end{aligned} \tag{1}$$

Up until this point, our approach is identical to that of Rossi, Gilula and Allenby (2001). Our approach now deviates as we assume that the cut-points are known and are given by $\lambda_0 = -\infty$, $\lambda_1 = 1.5$, $\lambda_2 = 2.5$, $\lambda_3 = 3.5$, $\lambda_4 = 4.5$ and $\lambda_5 = \infty$. We suggest that the chosen cut-points correspond to the way that respondents actually think. When asked to supply information on a five-point scale, we hypothesize that respondents make assessments on the continuum where the values $1.0, \ldots, 5.0$ have precise meaning. The respondents then implicitly round the continuous score to the nearest of the five integers. Although our methodology can be modified using unknown cut-points, the estimation of cut-points introduces difficulties involving nonidentifiability. Rossi, Gilula and Allenby (2001) address nonidentifiability and parsimony by imposing numerous constraints on the cut-points.

It is interesting to compare our rationale for the $Y_{ij} \rightarrow X_{ij}$ transformation with the range-frequency model proposed by Parducci (1965). The *range* principle suggests that a respondent uses extreme stimuli to fix the interpretation of endpoints on a discrete scale, and these endpoints provide reference for intermediate scale values. The principle is consistent with our transformation rationale as rounding is a subsequent step to marking latent variables on a continuum. On the other hand, the *frequency* principle appears to be violated as there is no reason to expect constant frequencies between scales values. This departure may be expected on the grounds of a *reference point effect* where Likert scale values, for example, have specific meanings. The frequency-range model and various departures from the model are discussed in Tourangeau et al. (2000).

Using the notation $Y_i = (Y_{i1}, \ldots, Y_{im})'$, Rossi, Gilula and Allenby (2001) consider

$$Y_i \sim \text{Normal}(\mu + \tau_i \underline{1},\ \sigma_i^2 \Sigma) \tag{2}$$

for $i = 1, \ldots, n$ where $\tau_i$ and $\sigma_i$ are respondent-specific parameters used to address *scale usage heterogeneity*. For example, a large $\tau_i$ and small $\sigma_i > 0$ characterize a respondent who uses the top end of the scale. Further, the model (2) implies a standardized response $(Y_{ij} - \mu_j - \tau_i)/\sigma_i$ through which the correlation between survey questions may be assessed. A consequence of the model is that correlation inferences between survey questions may differ considerably when scale usage characteristics are considered.

Although (2) contains many of the features we desire, it cannot, for example, adequately model an individual whose responses are mostly intermediate values such as 2's and 4's. We instead consider a structure that has similarities to (2). We propose

$$Y_i \sim \text{Normal}(b_i(\mu + a_i \underline{1} - 3\underline{1}) + 3\underline{1},\ b_i^2 \Sigma) \tag{3}$$

where we adjust for personalities via a "pure" or standardized score for the $i$th individual given by $Z_i = (Z_{i1}, \ldots, Z_{im})' \sim \text{Normal}(\mu,\ \Sigma)$ such that

$$Y_{ij} = b_i(Z_{ij} + a_i - 3) + 3 \tag{4}$$

for $i = 1, \ldots, n$, $j = 1, \ldots, m$.

It is (4) that provides an interpretation for the latent responses $Z_i$ and $Y_i$, and for the parameters $a_i$ and $b_i$ corresponding to the $i$th individual. We observe that $Z_i$ is a standardized latent score which is independent and identically distributed across respondents. The vector $\mu$ corresponds to the mean response of standardized scores over the population of respondents, and the matrix $\Sigma$ describes the variability of these scores and the correlation between survey questions. The latent score $Y_i$ is obtained from $Z_i$ via (4) where $Y_i$ includes the personality characteristics $(a_i, b_i)$ of the $i$th respondent. Unlike the $Z_i$, we note that the $Y_i$ in (3) are not identically distributed. Therefore, the learning of $(a_i, b_i)$ can be thought of as a denoising method where the pure response $Z_i$ is derived from the noisy $Y_i$ which includes personality traits.

For an interpretation of the *disposition* parameter $a_i \in \mathcal{R}$ in (4), it is initially helpful to consider $a_i$ conditional on $b_i = 1$. In this case, when $a_i = 0$, the $i$th respondent has a neutral disposition and the latent response $Y_{ij}$ is equal to the standardized score $Z_{ij}$. When $a_i > 0$ ($a_i < 0$), the $i$th respondent has a positive (negative) attitude since $Z_{ij}$ is adjusted by $a_i$ to give $Y_{ij}$.

For an interpretation of the *extremism* parameter $b_i > 0$ in (4), it is helpful to consider $b_i$ conditional on $a_i = 0$. In this case, when $b_i > 1$, the amount by which $Z_{ij}$ exceeds 3.0 is magnified and is added to 3.0 and gives a more extreme result towards the tails on the five-point scale. When $0 \leq b_i < 1$, the extremism parameter has the effect of pulling the latent response $Y_{ij}$ closer to the middle. A respondent whose $b_i \approx 0$ might be described as moderate and we impose the constraint $b_i > 0$ to avoid nonidentifiability. Note that the parameter $\sigma_i$ in (2) addresses variability which is somewhat different from our concept of extremism.

To provide a little more clarity, when $Z_{ij} + a_i - 3 > 0$, the $i$th respondent is positively inclined towards survey question $j$. When $Z_{ij} + a_i - 3 < 0$, the $i$-th respondent is negatively inclined towards survey question $j$. The quantity $Z_{ij} + a_i - 3$ is then scaled by $b_i$ to account for extremism on the part of the $i$-th respondent. The personality differential $b_i(Z_{ij} + a_i - 3)$ is then added to 3 to yield the latent variable $Y_{ij}$. Note that whereas a zero score for $b_i(Z_{ij} + a_i - 3)$ represents ambivalence (neither agree nor disagree in the Likert setting), $Y_{ij} = 3$ represents ambivalence in the latent variable. Having adjusted for respondent personalities, we are interested in the average response $\mu$ for the $m$ questions and the corresponding correlation structure $\Sigma$. We recognize that not all individuals share the same temperment. The $i$-th respondent is characterized by the parameters $a_i$ and $b_i$ where $a_i$ is the disposition parameter and $b_i$ is the extremism parameter.

As the proposed approach is Bayesian, prior distributions are required for the model parameters in (3). Specifically, we assign moderately diffuse priors

$$\begin{aligned} \Sigma^{-1} &\sim \text{Wishart}_m(I, m) \\ \mu_j &\sim \text{Uniform}(0, 6) \end{aligned}$$

where the components of $\mu = (\mu_1, \ldots, \mu_m)'$ are apriori independent. The Wishart distribution is the standard and conjugate prior distribution for the inverse covariance matrix in normal models (Bernardo and Smith 1994) where the identity matrix and degrees of freedom parameter $m$ are convenient choices in the absence of subjective prior information. Regarding the parameters $\mu_j$, although it is tempting to assign flat improper priors, our rationale for the Uniform$(0, 6)$ prior distribution is based on the observed response $X_{ij}$ constrained to the five-point scale. It is thought that $X_{ij}$ represents the rounded score of the continuous latent variable $Y_{ij}$ whose mean is $\mu_j$ when $b_i = 1$ and $a_i = 0$. For the personality parameters $a_i$ and $b_i$, the prior assignment is based on the supposition that there are classes of personalities. We therefore consider the Dirichlet process

$$\begin{aligned} (a_i, b_i)' &\overset{\text{iid}}{\sim} G \\ G &\sim \text{DP}(\alpha, \text{tr-Normal}_2\left(\mu_G, \Sigma_G\right)) \end{aligned} \tag{5}$$

for $i = 1, \ldots, n$. The specification in (5) states that $(a_i, b_i)$ arises from a distribution $G$ but $G$ itself arises from a distribution of distributions known as the Dirichlet process. The Dirichlet process in (5) consists of the concentration parameter $\alpha$ and baseline distribution tr-Normal$_2\left(\mu_G, \Sigma_G\right))$ where tr-Normal refers to the truncated bivariate Normal whose second component $b_i$ is constrained to be positive. The baseline distribution serves as an initial guess of the distribution of $(a_i, b_i)$ and the concentration parameter determines our confidence in the baseline distribution with large values of $\alpha > 0$ corresponding to greater degrees of belief. Prior distributions can be assigned to the hyperparameters in (5). Our analyses involving course evaluation surveys on a five-point scale give sensible results with $\alpha \sim$ Uniform$(0.4, 10)$ (Ohlssen, Sharples and Spiegelhalter, 2007), $\mu_G = (0, 1)'$ and $\Sigma_G = (\sigma_{ij})$ where $\sigma_{11} = 1.0$, $\sigma_{22} = 0.5$ and $\sigma_{ij} = 0$ for $i \neq j$. Note that the choice of 1.0 and 0.5 are sufficiently diffuse in the range of parameters $a_i$ and $b_i$. The key aspect of the Dirichlet process in our application is that the personality parameters $(a_i, b_i)$ have support on a discrete space and this enables the clustering of personality types. An advantage of the Dirichlet process approach is that clustering is implicitly carried out in the framework of the model and the number of component clusters need not be specified in advance. Once a theoretical curiousity, the Dirichlet process and its extensions are finding diverse application areas in nonparametric modelling (e.g. Qi, Paisley and Carin 2007, Dunson and Gelfand 2009, Gill and Casella 2009). The nonparametric prior specification in our model and the associated clustering of subjects provides another essential difference between our approach and that of Rossi, Gilula and Allenby (2001).

# 3 COMPUTATION

The model described in section 2 is generally referred to as a Dirichlet process mixture model, and various Markov chain methodologies have been developed to facilitate sampling-based analyses (Neal 2000). However, these algorithms require considerable sophistication on the part of the programmer.

A goal in this paper is to simplify the programming aspect of the analysis by carrying out computations in WinBUGS. The basic idea behind WinBUGS is that the programmer need only specify the statistical model, the prior and the data. The Markov chain calculations are done in the background whereby the user is then supplied with Markov chain output. Markov chain output is then conveniently averaged to give approximations of posterior means.

To implement the analysis of our model in WinBUGS, we make use of the constructive definition of the Dirichlet process given by Sethuraman (1994). The definition is known as the stick breaking representation, and in the context of our problem, it is given as follows: Generate a set of iid atoms $(a_i^*, b_i^*)$ from tr-Normal$_2\,(\mu_G, \Sigma_G)$ and generate a set of weights $w_i = y_i \prod_{j=1}^{i-1}(1 - y_j)$ where the $y_i$ are iid with $y_i \sim \text{Beta}(1, \alpha)$ for $i = 1, \ldots, \infty$. Then

$$G = \sum_{i=1}^{\infty} w_i \delta_{(a_i^*, b_i^*)} \tag{6}$$

where $\delta_{(a_i^*, b_i^*)}$ is the point mass at $(a_i^*, b_i^*)$.

For programming in WinBUGS, the Sethurman (1994) construction is most useful as it allows us to approximately specify the prior. We see that the stick breaking mechanism creates smaller and smaller weights $w_i$. This suggests that at a certain point we can truncate the sum (6) and obtain a reasonable approximation to $G$ (Muliere and Tardella 1998). Ishwaran and Zarepour (2002) suggest that the number of truncation points be $n$ when the sample size is small and $\sqrt{n}$ when the sample size is large. The stick breaking construction clearly shows that a generated $G$ is a discrete probability distribution which implies that there is non-negligible probability that $(a_i, b_i)$'s generated from the same $G$ have the same value. This facilitates the clustering of personalities in ordinal survey data. We note that the original definition of the Dirichlet process (Ferguson 1973) does not provide a WinBUGS-tractable expression for the prior.

# 4 EXAMPLES

## 4.1 Course Evaluation Survey Data

The proposed model is fit to data obtained from teaching and course evaluations in the Department of Statistics and Actuarial Science at Simon Fraser University (SFU). The standard questionnaire at SFU contains $m = 15$ questions with responses on a five-point scale ranging from 1 (a very negative response) to 5 (a very positive response) where the specific interpretation of responses are question dependent. The survey questions are given as follows:

1. The course text or supplementary material was
2. I would rate this course as
3. The assignments and lectures were
4. The assignments and exams were on the whole

5. The marking scheme was on the whole

6. How informative were the lectures

7. The Instructor's organization and preparation were

8. The Instructor's ability to communicate material was

9. The Instructor's interest in the course content appeared to be

10. The Instructor's feedback on my work was

11. Questions during class were encouraged

12. Was the Instructor accessible for extra help

13. Was the Instructor responsive to complaints/suggestions

14. Overall, the Instructor's attitude towards students was

15. I would rate the Instructor's teaching ability as

Data were collected from $n = 75$ students pertaining to an introductory Statistics course. Posterior means and standard deviations corresponding to the parameter $\mu$ are given in Table 1. These are based on a MCMC simulation using WinBUGS with a burn-in period of 1000 iterations followed by 4000 iterations, taking roughly 2 hours of computation on a personal computer. The WinBUGS code is provided in the Appendix. The highest posterior mean was recorded for the 9th question which asked about "the Instructor's interest in the course material". The smallest mean was recorded for the 10th question which asked about "the Instructor's feedback on work". These results are consistent with past surveys taken in the same course with the same Instructor. In particular, the Instructor does not grade assignments and this yields some criticisms from the students. Note that the posterior standard deviations are sufficiently small such that we can sensibly discuss the posterior means.

To investigate the clustering effect, we recorded the number of clusters in each of the

Markov chain simulations. The resulting histogram is given in Figure 1. We observe that there are quite a few clusters, and there is considerable uncertainty about the number of clusters. More specifically, the number of clusters appears roughly uniform between 3 and 18, with approximately 10 clusters on average. This implies that there is a substantial number of personality types amongst the $n = 75$ students. The clustering effect is corroborated in Figure 2 where we provide a plot of the posterior means of the $(a_i, b_i)$ pairs. By looking closely along both vertical and horizontal strips, there are approximately 10 classes of personalities, some of which do not differ greatly. In Figure 2, the clustering is more difficult to distinguish with respect to the disposition parameter $a$, suggesting more variability in $a$ than in $b$. We observe that roughly 50% of the $a_i$'s are greater than 0.0, and roughly 50% of the $b_i$'s are greater than 1.0.

For small values of $b$, the corresponding students tend to have responses which are often the same, and this may be due to a desire to finish the questionnaire as quickly as possible. With $b = 1.0$, the interpretation is that these students do not distort their responses in an inflationary/deflationary sense, and there are roughly 20 students of this type. For large values of $b$, the corresponding students inflate their responses; they make harsher decisions near both ends of the scale (1's and 5's).

Note that one of the respondents provided a score of 5.0 for all $m = 15$ questions. It turns out that the corresponding posterior mean of $(a_i, b_i)$ for this student was $(0.36, 1.09)$. Based on an average posterior response $\bar{\mu} = 4.1$, this student's mean latent Y-score is $1.09(\bar{\mu} + 0.36 - 3) + 3 = 4.59$ which rounds to a respondent X-score of 5.0 according to the cut-point model (1). This provides some evidence that the $(a_i, b_i)$ parameters are estimated sensibly. For this student, we note that the variance $1.09\Sigma$ of the latent response $Y_i$ which includes personality traits exceeds the variance $\Sigma$ of the standardized response $Z_i$. This is because the student in question is "extreme", and has the capacity for extreme responses of 0's and 5's. This case highlights a distinction between our notion of extremism and variability.

As another example of the adjustment made for personalities, the smallest posterior mean for the disposition parameter corresponds to $a_i = -0.30$ which was recorded for a student with an average observed response $\bar{X}_i = 2.06$ over all $m = 15$ questions. This student has a corresponding extremism parameter $b = 1.01$. The question arises as to whether 2.06 is a measurement that should be taken at face value when the average response and standard deviation over all students are 3.89 and 1.02 respectively. It appears that 2.06 is an extreme score lying 1.8 standard deviations from the mean. However, when we adjust for the personality of the student via (4), we obtain $Y_i = 1.01(Z_i - 0.30\underline{1} - 3\underline{1}) + 3\underline{1} \approx Z_i - 0.3\underline{1}$. This implies that the standardized but latent response $Z_i$ is larger than $Y_i$. The student has a negative disposition, and when we account for the negative disposition, the de-noised score $Z_i$ is not as extreme as the raw data $X_i$.

It is also instructive to look at the posterior mean of the variance-covariance matrix $\Sigma$ which describes the relationships amongst the $m = 15$ survey questions. The largest correlation 0.63 occurred between survey questions 14 and 15. This is consistent with our intuition and personal teaching experience whereby students think highly of their instructors when they believe that their instructors care about them. The second highest correlation 0.57 occurred between survey questions 6 and 7 which is also believable from the view that learning is best achieved when material is clearly presented. However, we emphasize that the elimination of questions on the basis of redundancy should not be done solely on the basis of high correlations. In addition to high correlations, we should also have similar posterior means. With the estimated posterior means $\mu_{14} = 4.57$ and $\mu_{15} = 4.52$, SFU may feel comfortable in dropping either question 14 or question 15 from the survey. Furthermore, we note that there were no negative posterior correlations and

Table 1: Posterior means and posterior standard deviations for the SFU survey data.

| Parameter | Posterior Mean | Posterior SD |
|---|---|---|
| $\mu_1$ | 3.69 | 0.19 |
| $\mu_2$ | 3.53 | 0.15 |
| $\mu_3$ | 4.04 | 0.18 |
| $\mu_4$ | 3.45 | 0.17 |
| $\mu_5$ | 3.85 | 0.17 |
| $\mu_6$ | 4.54 | 0.19 |
| $\mu_7$ | 4.33 | 0.18 |
| $\mu_8$ | 4.41 | 0.17 |
| $\mu_9$ | 4.78 | 0.15 |
| $\mu_{10}$ | 3.23 | 0.17 |
| $\mu_{11}$ | 4.51 | 0.19 |
| $\mu_{12}$ | 4.01 | 0.18 |
| $\mu_{13}$ | 4.11 | 0.17 |
| $\mu_{14}$ | 4.57 | 0.19 |
| $\mu_{15}$ | 4.52 | 0.18 |
| $\bar{\mu}$ | 4.10 | |

the minimum correlation 0.11 occurred between question 1 and question 13. Our intuition accordingly suggests that these two questions are independent. For comparison, we have also calculated the sample correlation matrix based on the raw scores $X$. The values align with the posterior mean of $\Sigma$. For example, the smallest sample correlation is 0.10 and this is observed between question 1 and question 13. The largest sample correlation is 0.78 and this occurs between questions 14 and 15.

It is good statistical practice to look at various plots related to the MCMC simulation. Trace plots for the parameters appear to stabilize immediately and hence provide no indication of lack of convergence in the Markov chain. Furthermore, autocorrelation plots appear to dampen quickly. This provides added evidence of the convergence of the Markov chain and also suggests that it may be appropriate to average Markov chain output as though the variates are independent. In addition to the diagnostics described, multiple chains were generated to provide further assurance of the reliability of the methods. For example, the Brooks-Gelman-Rubin statistic (Brooks and Gelman 1997) gave no indication of lack of convergence.

We now consider the analysis of the SFU survey data using the methodology of RGA (Rossi, Gilula and Allenby 2001). Whereas our model uses known cut-points which convert the latent variable $Y_{ij}$ to the observed $X_{ij}$, RGA have cut-points that are determined via constraints and a single unknown parameter $e$. For the RGA analysis, $\lambda_i = c + di + ei^2$, $i = 1, \ldots, 4$, and the constraints $\sum_{i=1}^4 \lambda_i = 12$ and $\sum_{i=1}^4 \lambda_i^2 = 41$ were imposed such that the cut-points are apriori centred about the known cut-points in our model where $e \sim \text{Uniform}(-0.2, 0.2)$.

Fitting the RGA model, we obtained posterior means $e = -0.003$, $\lambda_1 = 1.50$, $\lambda_2 = 2.51$, $\lambda_3 = 3.51$ and $\lambda_4 = 4.48$ where we observe that the RGA cut-points are very close to the fixed cut-points used in our model. To compare the fit of the RGA model with our model using the SFU survey data, we calculated the posterior mean of the diagnostic $D = \sum(y_{ij} - \beta_{ij})^2$ where $\beta_{ij}$ denotes the mean of $y_{ij}$ and the summation is taken over all pairs $(i, j)$ where $x_{ij} \neq 1$ and $x_{ij} \neq 5$ (see (1)). The restricted summation is imposed since the RGA model does not impose lower and upper values for $y_{ij}$, and consequently small/large posterior variates $y_{ij}$ greatly inflate the diagnostic $D$. The diagnostic $D$ is in the spirit of deviances (McCullagh and Nelder 1989) where $y_{ij}$ denotes the underlying latent score in both the RGA model and in our model. In the RGA model (2), $\beta_{ij} = \mu_j + \tau_i$, and in our model (3), $\beta_{ij} = b_i(\mu_j + a_i - 3) + 3$. Whereas the RGA model gave $D = 936$, our model gave $D = 891$. In both the RGA model and in our model, $\mu$ denotes the vector of standardized scores.

For the sake of comparison, the posterior means and posterior standard deviations of $\mu_{15}$ (the standardized score for the 15th survey question) are 4.38(0.15) and 4.52(0.18) for the RGA model and for our model respectively. We also consider a particular student; one who recorded low values (six 1's, two 2's and seven 3's) on the course evaluation survey. This student has posterior means $a = -0.30$ and $b = 1.01$ indicating that the student has a negative disposition but typical extremism. In the RGA model, the student has posterior characteristics $\tau = -0.13$, and $\sigma = 0.99$. Although $(a, b)$ and $(\tau, \sigma)$ are not directly comparable, it seems that both models captured the essence of this student. Therefore, from various perspectives, the RGA model and our model give comparable results in this example.

To investigate an aspect of the internal consistency of the methodology, we collapse the five-point scale to a three-point scale. The original data matrix $X$ is recoded so that negative scores (1's and 2's) are coded as 1's, moderate scores (3's) are coded as 2's, and positive scores (4's and 5's) are coded as 3's. Accordingly, we set cut-points $\lambda_0 = -\infty$, $\lambda_1 = 1.5$, $\lambda_2 = 2.5$ and $\lambda_3 = \infty$. Following (4), a standardized score $Z_i = (Z_{i1}, \ldots, Z_{im})'$ is defined via $Y_{ij} = b_i(Z_{ij} + a_i - 2) + 2$. And in a similar fashion to the model based on the five-point scale, we consider the prior $\mu_j \sim \text{Uniform}(0, 3)$. To get a sense of agreement between the model based on the five-point scale and the collapsed model based on the three-point scale, we calculate the difference $d_{ai} = a_{5i} - a_{3i}$ where $a_{5i}$ and $a_{3i}$ are the corresponding posterior means of the disposition parameter for the $i$th subject, $i = 1, \ldots, n$. We then calculate the sample standard deviation $s_a = 0.16$. Similarly, we calculate the sample standard deviation $s_b = 0.09$ corresponding to the extremism parameter. Referring to Figure 2, the sample standard deviations suggest reasonable agreement between the model based on the five-point scale and the collapsed model based on the three-point scale. Of course, we should not expect perfect agreement, especially on the $b$-parameter (extremism), since it is difficult to be characterized as extreme when there are only three possible responses.

## 4.2 Simulated Data

Several simulation studies were carried out to investigate the model. We report on one such simulation. A dataset corresponding to $n = 150$ subjects with $m = 10$ questions was simulated using R code. In this example, the mean vector $\mu = (3, 3, 3, 3, 3, 4, 4, 4, 4, 4)'$ and variance covariance matrix $\Sigma = (\sigma_{ij})$ with $\sigma_{ii} = 4$ and $\sigma_{ij} = 2$ for $i \neq j$ were used to generate the latent matrix $Z$. The personality parameters $a_i$ and $b_i$ were set according to $(a_i, b_i) = (0.0, 1.0)$ for the first 75 subjects and $(a_i, b_i) = (0.2, 0.8)$ for the remaining 75 subjects. Having generated $Z$ as described, we then obtained $Y$ via (4) and then obtained the observed data matrix $X$ using the cut-point model (1).

The model was fit using WinBUGS software where 1000 iterations were used for burn-in. The posterior statistics in Table 2 were based on 4000 iterations. We observe that the posterior means of the mean vector are in rough agreement with the true $\mu$. The posterior means of $\Sigma$ are also consistent with the underlying values. The level of agreement is high because we have many subjects ($n = 150$) relative to questions ($m = 10$). The level of agreement improved as we increased the number of respondents $n$.

In another simulation, we considered large $m$ (number of survey questions) relative to $n$ (number of subjects). As anticipated, the posterior means of the personality parameters $(a_i, b_i)$ were in agreement with the true model parameters, $i = 1, \ldots, n$.

Table 2: Posterior means and posterior standard deviations for the simulated data.

| Parameter | Posterior Mean | Posterior SD |
|---|---|---|
| $\mu_1$ | 3.08 | 0.15 |
| $\mu_2$ | 3.00 | 0.12 |
| $\mu_3$ | 3.17 | 0.16 |
| $\mu_4$ | 3.14 | 0.17 |
| $\mu_5$ | 2.96 | 0.18 |
| $\mu_6$ | 4.10 | 0.14 |
| $\mu_7$ | 4.19 | 0.17 |
| $\mu_8$ | 3.98 | 0.19 |
| $\mu_9$ | 3.92 | 0.15 |
| $\mu_{10}$ | 4.10 | 0.20 |

# 5 GOODNESS-OF-FIT

In Bayesian statistics, there is no consensus on the "correct" approach to the assessment of goodness-of fit. When Bayesian model assessment is considered, it appears that the prominent modern approaches are based on the posterior predictive distribution (Gelman, Meng and Stern 1996). These approaches rely on sampling future variates $y$ from the posterior predictive density

$$f(y \mid x) = \int f(y \mid \theta)\ \pi(\theta \mid x)\ d\theta \tag{7}$$

where $x$ is the observed data, $f(y \mid \theta)$ is the sampling density and $\pi(\theta \mid x)$ is the posterior density. In MCMC simulations, approximate sampling from (7) proceeds by sampling $y_i$ from $f(y \mid \theta^{(i)})$ where $\theta^{(i)}$ is the $i$th realization of $\theta$ from the Markov chain. Model assessment then involves a comparison of the future values $y_i$ versus the observed data $x$. One such comparison involves the calculation of posterior predictive p-values (Meng

1994). A major difficulty with posterior predictive methods concerns double use of the data (Evans 2007). Specifically, the observed data $x$ is used both to fit the model giving rise to the posterior density $\pi(\theta \mid x)$ and then is used in the comparison of $y_i$ versus $x$. For this reason, some authors prefer a cross-validatory approach (Gelfand, Dey and Chang 1992) where the data $x = (x_1, x_2)$ are split such that $x_1$ is used for fitting and $x_2$ is used for validation.

We take the view that in assessing a Bayesian model, the entire model ought to be under consideration, and the entire model consists of both the sampling model of the data and the prior. We also want a methodology that does not suffer from double use of the data. For the models proposed here, we recommend an approach that is similar to the posterior predictive methods but instead samples "model variates" $y$ from the prior predictive density

$$f(y) = \int f(y \mid \theta)\ \pi(\theta)\ d\theta \tag{8}$$

where $\pi(\theta)$ is a proper prior density. This approach was advocated by Box (1980) before simulation methods were common. It is not difficult to write R code to simulate $y_1, \ldots, y_N$ from the prior predictive density in (8). It is then a matter of deciding how to compare the $y_i$'s against the observed data matrix $X$. In our application, the data are high dimensional, and we advocate a comparison of "features" that are of direct interest. This is an intuitive and simple approach which is not part of current statistical practice. For example, one might compare observed subject means $\bar{X}_i = \sum_{j=1}^{m} X_{ij}/m$ with subject means generated from the prior predictive simulation. A simple comparison of these vectors can be easily carried out through the calculation of Euclidean distances. Naturally, as the priors become more diffuse, it becomes less likely to find evidence of model inadequacy. We do not view this as a failing of the methodology. Rather, if you really want to detect departures from a model, it is necessary that you have strong prior opinion concerning your model.

To provide a more stringent test, we consider a modification of our model where subjective priors $\mu_j \sim \text{Uniform}(2,5)$ and $\Sigma_G = 0.01I$ are introduced. We assess goodness-of fit on the SFU data discussed in section 4. With $N$ simulated vectors from the prior predictive distribution, there are $\binom{N+1}{2}$ Euclidean distances of interest; $N$ of these distances are between the observed mean vector and the simulated vectors, and the remaining $\binom{N}{2}$ distances correspond to distances between simulated vectors. These distances are displayed in a histogram with the $N = 20$ distances highlighted in Figure 3. Since these distances appear typical, there is no evidence of lack of fit. In fact, we observe that most of the Euclidean distances involving observed data lie on the left side of the histogram. This suggests that the most extreme variates arose from the prior-predictive distribution. Clearly, graphical displays for alternative features can also be produced.

Another approach to Bayesian goodness-of-fit which appears promising in the context of the proposed model is due to Johnson (2007). Let $\theta$ consist of all model parameters, let $X_i$ be the vector of discrete responses for the $i$th respondent and let $\beta_i = b_i(\mu + a_i\underline{1} - 3\underline{1}) + 3\underline{1}$ denote the mean of the corresponding latent variable $Y_i$, $i = 1, \ldots, n$. Under the "true" $\theta$, we then note that $S(X_i, \theta) = (Y_i - \beta_i)'\Sigma^{-1}(Y_i - \beta_i)/b_i^2$ is distributed as a Chi-square variable with $m = 15$ degrees of freedom. Following Johnson (2007), $S(X_i, \theta)$ is pivotal in the sense that its conditional distribution does not depend on $\theta$ and there are $n = 75$ values of $S(X_i, \theta)$ that can be calculated for a given $\theta$. For a single sampled value $\theta$ from the MCMC simulation, Figure 4 provides a plot of the ordered values of $S(X_i, \theta)$ versus the theoretical Chi-square quantiles. The plotted points appear to be roughly scattered about the line $y = x$ and hence provide no strong indication of lack of fit.

# 6 DISCUSSION

We have developed a Bayesian latent variable model to analyze ordinal response survey data. We have also facilitated a clustering mechanism based on personalities. Most importantly, clustering takes place as a consequence of Dirichlet process modelling of the personality parameters. In a WinBUGS programming environment, the model is succinctly formulated, and is not complicated by latent variables and missing data.

Our model identifies areas where performance has been poor or exceptional in a ordinal survey data by investigating standardized parameters. It also allows us to check whether some questions in a survey are redundant. A goodness-of-fit procedure is advocated that is based on comparing prior-predictive output versus observed data. The approach is intuitive and is flexible in the sense that one can investigate features which are relevant to the particular model. Future enhancements may be considered such as including subject covariates and handling longitudinal data structures.

One of the assumptions in our model concerns the use of fixed cut-points in transforming the underlying continuous latent responses $Y_{ij}$ to the observed discrete responses $X_{ij}$. Although it may have been preferable to allow variable cut-points, we were unable to implement the generalization. Issues of non-identifiablility and model complexity lead to Markov chains which did not achieve practical convergence.

# 7 REFERENCES

Agresti, A. (2010). *Analysis of Ordinal Categorical Data, Second Edition*, Wiley: New York.

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

Box, G. E. (1980). "Sampling and Bayes' inference in scientific modelling and robustness" (with discussion), *Journal of the Royal Statistical Society, Series A*, 143, 383–430.

Brooks, S. P. and Gelman, A. (1997). "Alternative methods for monitoring convergence of iterative simulations", *Computational and Graphical Statistics*, 7, 434–455.

Cowen, L., Manion, L. and Morrison, K. (2007). *Research Methods in Education, Sixth Edition*, Routledge: New York.

Dawes, J. (2008). "Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales", *International Journal of Market Research*, 50, 61-77.

Dolnicar, S. and Grun, B. (2007). "Cross-cultural differences in survey response patterns", *International Marketing Review*, 24, 127-143.

Dunson, D.B. and Gelfand, A.E. (2009). "Bayesian nonparametric functional data analysis through density estimation", *Biometrika*, 96, 149-162.

Emons, W.H.M. (2008). "Nonparametric person-fit analysis of polytomous item scores", *Applied Psychologial Measurement*, 32, 224-247.

Evans, M. (2007). Comment on "Bayesian checking of the second levels of hierarchical models" by Bayarri and Castellanos, *Statistical Science*, 22, 344-348.

Fanning, E. (2005). "Formatting a paper-based survey questionnaire: best practices", *Practical Assessment Research & Evaluation*, online: http://pareonline.net/pdf/v10n12.pdf.

Ferguson, T.S. (1973). "A Bayesian analysis of some nonparametric problems", *Annals of Statistics*, 1, 209-230.

Forrest, M. and Andersen, B. (1986). "Ordinal scale and statistics in medical research", *British Medical Journal*, 292, 537-538.

Gelfand, A. E., Dey, D. K. and Chang, H. (1992). "Model determination using predictive distributions with implementation via sampling-based methods" (with discussion), In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith, editors), Oxford: Oxford University Press, 147–167.

Gelman, A., Meng, X. L. and Stern, H. S. (1996). "Posterior predictive assessment of model fitness via realized discrepancies", *Statistica Sinica*, 6, 733–807.

Gill, J. and Casella, G. (2009). "Nonparametric priors for ordinal Bayesian social science models: specification and estimation", *Journal of the American Statistical Association*, 104, 453-464.

Goodmann, L.A. (1979). “Simple models for the analysis of association in cross-classifications having ordered categories”, *Journal of the American Statistical Association*, 74, 537-552.

Ishwaran, H. and Zarepour, M. (2002). “Dirichlet prior sieves in finite normal mixtures”, *Statistica Sinica*, 12, 941-963.

Javaras, K.N. and Ripley, B.D. (2007). “ An ‘unfolding’ latent variable model for Likert attitude data: Drawing inferences adjusted for response style”, *Journal of the American Statistical Association*, 102, 454-463.

Johnson, T.R. (2003). “On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style”, *Psychometrika*, 68, 563-583.

Johnson, V.E. (1996). “On Bayesian analysis of multirater ordinal data: An application to automated essay grading”, *Journal of the American Statistical Association*, 91, 42-51.

Johnson, V.E. (2007). “Bayesian model assessment using pivotal quantities”, *Bayesian Analysis*, 2, 719-734.

Kottas, A., Mueller, P. and Quintana, F. (2005). “Nonparametric Bayesian Modeling for Multivariate Ordinal Data”, *Journal of Computational and Graphical Statistics*, 14, 610-625.

Liu, I. and Agresti, A. (2005). “The analysis of ordered categorical data: An overview and a survey of recent developments”, *Test*, 14, 1-73.

McCullagh, P. (1980). “Regression models for ordinal data (with discussion)”, *Journal of the Royal Statistical Society*, Series B, 42, 109-142.

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

Meng, X.L. (1994). “Posterior predictive p-values”, *The Annals of Statistics*, 22, 1142–1160.

Muliere, P. and Tardella, L. (1998). “Approximating distributions of random functionals of Ferguson-Dirichlet priors”, *Canadian Journal of Statistics*, 26, 283-297.

Neal, R.M. (2000). “Markov chain sampling methods for Dirichlet process mixture models”, *Journal of Computational and Graphical Statistics*, 9, 249-265.

Ohlssen, D., Sharples, L.D. and Spiegelhalter, D.J. (2007). “Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons”, *Statistics in Medicine*, 26, 2088-2112.

Parducci, A. (1965). "Category judgment: A range-frequency model", *Psychological Review*, 72, 407-418.

Qi, Y., Paisley, J.W. and Carin, L. (2007). "Music analysis using hidden markov mixture models" *IEEE Transactions in Signal Processing*, 55, 5209-5224.

Rossi, P.E., Gilula, Z. and Allenby, G.M. (2001). "Overcoming scale usage heterogeneity", *Journal of the American Statistical Association*, 96, 20-31.

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors", *Statistica Sinica*, 4, 639-650.

Spiegelhalter, D. Thomas, A. and Best, N. (2003). *WinBUGS (Version 1.4) User Manual*, Cambridge: MRC Biostatistics Unit.

Tourangeau, R., Rips, L. and Rasinski, K. (2000). *The Psychology of Survey Response*, Cambridge University Press: Cambridge.

# 8 APPENDIX

We provide the WinBUGS code used in the analysis of the SFU survey data.

```
model
{
# cut point model as defined in (1)
alpha[1]<- -5; alpha[6]<-10; alpha[2]<- 1.5
alpha[3]<- 2.5; alpha[4]<-3.5; alpha[5]<-4.5

# multivariate normal structure as defined in (3) and (4)
for(i in 1:n) {for(j in 1:m)
{lo[i,j]<-((alpha[x[i,j]]-3)/b[i])+3 -a[i]
up[i,j]<-((alpha[x[i,j]+1]-3)/b[i])+3 -a[i]}}
for(i in 1:n) {z[i,1:m]~dmnorm(mu[],G[,])I(lo[i,],up[i,])}

# priors for mu and sigma
for(i in 1:m) {mu[i]~dunif(0,6)}
G[1:m,1:m]~dwish(R[,],m)
varcov[1:m,1:m]<-inverse(G[,])
for(j in 1:m) {cor[j,j]<-varcov[j,j]}
for(i in 1:m-1) {for(j in i+1:m)
{cor[i,j]<-varcov[i,j]/(sqrt(varcov[i,i]*varcov[j,j])); cor[j,i]<-cor[i,j]}}

# DP Priors for a's and b's as in (6)
for(i in 1:n) {a[i]<-aa[i,1];b[i]<-(aa[i,2])}
for(j in 1:n) {for (kk in 1:2) {aa[j,kk]<-theta1[latent[j],kk]}}
for(i in 1:n) {latent[i]~dcat(pi[1:L1])} pi[1]<-r[1]
for(j in 2:(L1-1)) {log(pi[j])<-log(r[j])+sum(R1[j,1:j-1])
for(l in 1:j-1) {R1[j,l]<-log(1-r[l])}} pi[L1]<-1-sum(pi[1:(L1-1)])
for(j in 1:L1) {r[j]~dbeta(1,mm)}

# baseline distribution for DP as in (5)
for(i in 1:L1) {theta1[i,1:2]~dmnorm(zero[1:2],Sab[1:2,1:2])I(LB[],)}
zero[1]<-0; zero[2]<-1
Sab[1:2,1:2]~dwish(Omega[1:2,1:2],2); varcovab[1:2,1:2]<-inverse(Sab[,])
corab<-varcovab[1,2]/sqrt(varcovab[1,1]*varcovab[2,2])

# prior for concentration parameter
mm~dunif(0.4,10)  }
```

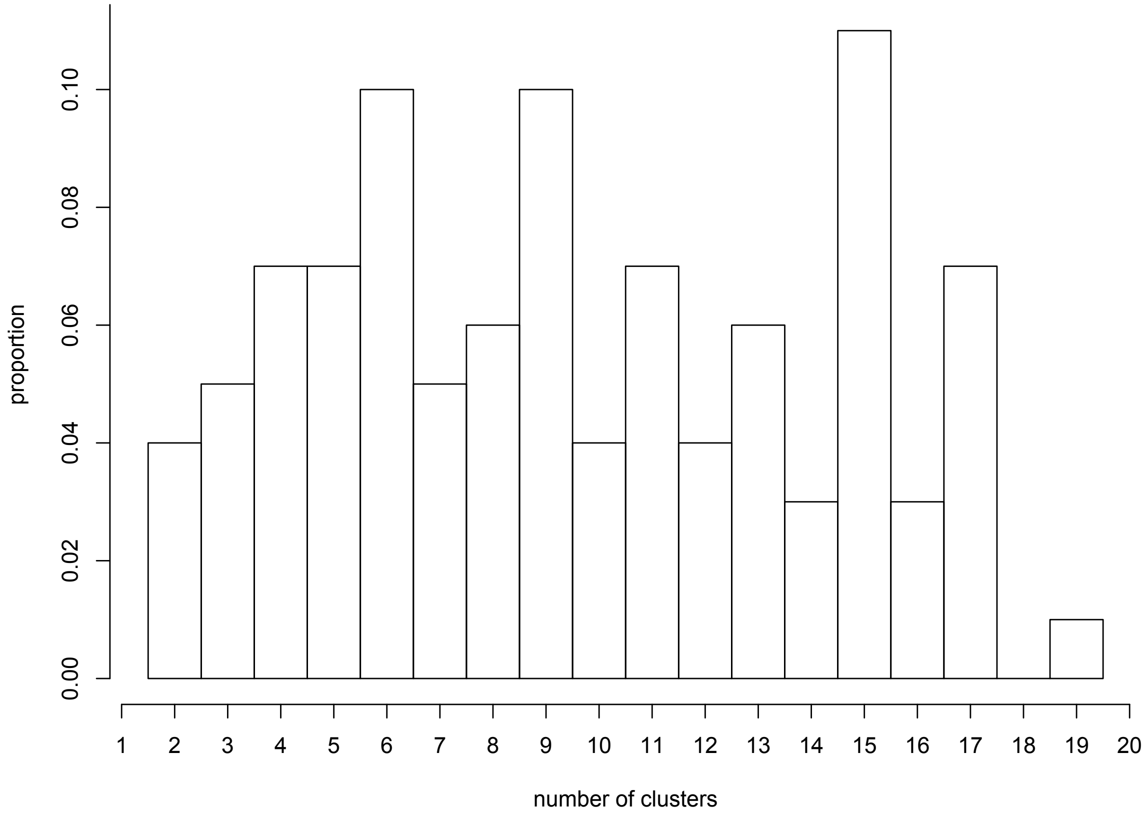

Figure 1: Histogram of the number of clusters for the SFU survey data.

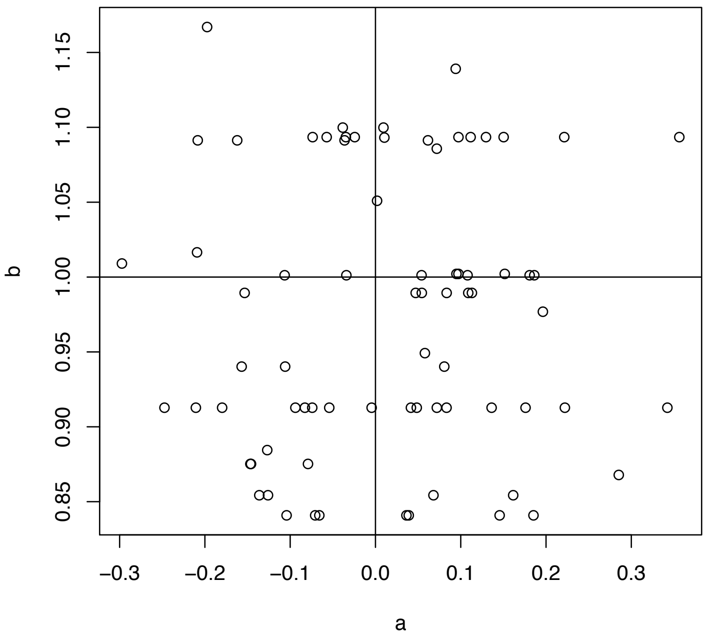

Figure 2: Plot of the posterior means of the personality parameters $(a_i, b_i)$ for the SFU survey data.

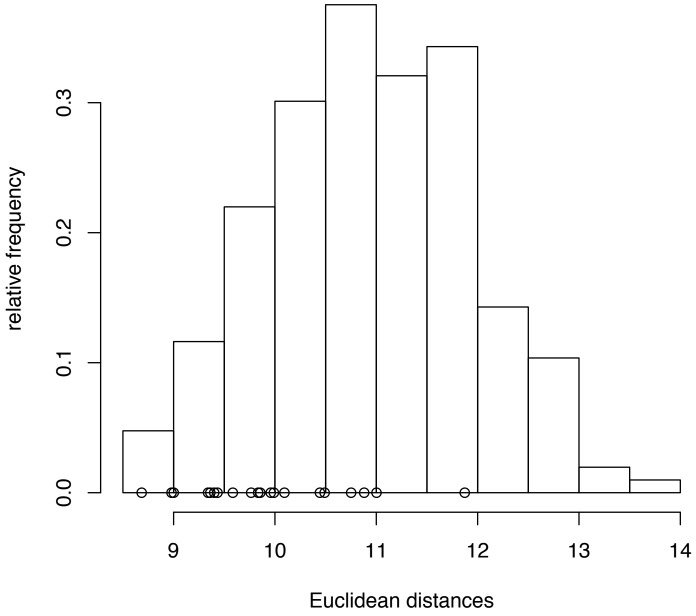

Figure 3: Histogram corresponding to the $\binom{N+1}{2}$= 210 Euclidean distances with respect to the prior-predictive check for the SFU survey data.

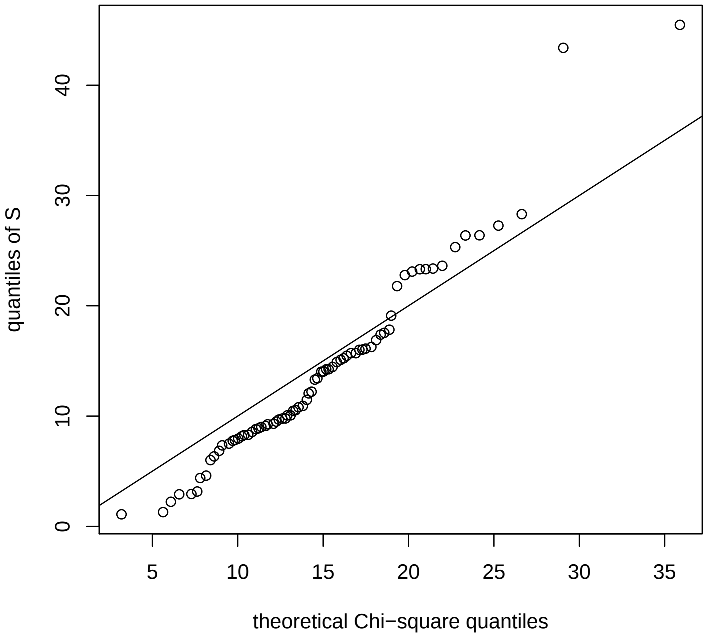

Figure 4: Q-Q plot used to investigate model fit as proposed by Johnson (2007) for the SFU survey data.