A Bayes Inference for Ordinal Response with Latent Variable Approach

Naijun Sha *† and Benard Owusu Dechi

Department of Mathematical Sciences, University of Texas at El Paso, El Paso, TX 79968, USA; bdechi@miners.utep.edu
* Correspondence: nsha@utep.edu; Tel.: +1-915-747-6844
† Current address: 500 W University Ave., El Paso, TX 79968, USA.

Received: 18 May 2019; Accepted: 14 June 2019; Published: 16 June 2019

Abstract: In this paper, we propose a Bayesian model for the analysis of categorical data with an ordered outcome. The method provides a latent variable approach with an informative prior transformed from a Dirichlet distribution for the boundary parameters. A simulation study is carried out to assess the performance of the methods under various settings of the data structure. Our method produces predictive accuracy over the conventional classification procedures. Real data are analyzed to demonstrate the efficiency of the proposed method.

Keywords: ordinal outcome; latent variable; Bayesian inference; Dirichlet prior; MCMC sampling

1. Introduction

Over the past few decades, modeling and predicting of ordinal outcomes have become essential to study in various fields especially in social and economic sciences where the natural ordering data very commonly appear. For example, socioeconomic status is typically broken into three levels (high, middle, and low) to describe the three places into which a family or an individual may fall. The level of education consists of high school, bachelor’s, master’s, and doctoral degrees. These levels or classes can be viewed as ordinal variables, but with no scale or magnitude available between each order or category. Many such ordinal data and numerous methods of analysis have been introduced and discussed by a number of researchers: [1–3], just to name a few. One well-known method is polytomous ordinal logistic regression (POLR) or the cumulative logit model initially proposed by [4], later called the proportional odds model [2] since the same proportionality constant applies for all cumulative logits. For a multinomial response variable \( Z \) with possible \( J \) ordered categorical outcomes and the associated \( p \)-dimensional vector of covariates \( x \), the cumulative probability for \( Z \) on \( x \) is given by:

\[
P(Z \leq j|x) = \frac{e^{(\alpha_j + x'\beta)}}{1 + e^{(\alpha_j + x'\beta)}}, \quad j = 1, 2, \ldots, J - 1, \tag{1}
\]
or the cumulative logit form as:

\[
\log \left( \frac{P(Z \leq j|x)}{P(Z > j|x)} \right) = \alpha_j + x'\beta, \quad j = 1, 2, \ldots, J - 1, \tag{2}
\]

where \( \alpha_j \) is an unknown intercept parameter associated with the \( j \)th category and \( \beta = (\beta_1, \beta_2, \ldots, \beta_p)' \) is the common vector of effect coefficients across the categories. The POLR models the cumulative probabilities \( P(Z \leq j) \) rather than the specific category probabilities \( P(Z = j) \) as in the nominal logistic regression. Implementing nominal analysis on the ordinal data will lead to underestimation.
of the variation and considerable loss of information [1]. Other relevant approaches dealing with ordinal outcomes include adjacent-category logit and sequential logit models. A number of researchers have proposed various inference processes for the logistic regression or other methods by classical approaches in [2,5] or from Bayesian perspectives in [6,7]. A comprehensive review about the analysis of ordered category data can be found in [1]. Another commonly-adopted way of modeling ordinal data is via an underlying continuous latent variable. This regards the observed ordinal response as a crude measurement of a continuous variable falling into an interval on the real line. The work in [3] applied a Bayesian analysis approach for a latent variable model to investigate the effect of a binary treatment on an ordinal outcome of interest. A fully-Bayesian method for modeling polychotomous ordinal categories was developed in [6] by using the data augmentation approach. Some issues arise with these approaches, one of which is the estimation of the cutpoint parameters of the interval boundary. Incorporating a vague prior on these parameters, the work in [6] proposed a probit model in a Bayesian framework, which leads to a slow convergence under large sample sizes because of inefficient sampling of the cutoff point parameters. The work in [7] suggested a hybrid Gibbs/Metropolis–Hasting (MH) sampling scheme, which updates these parameters jointly with other parameters. Although the approach reduces the high auto-correlations in the sampling, it may require the computation of joint cumulative probabilities of multivariate distributions, which is generally intractable, for computing the acceptance probability in MH samplers. To avoid these difficulties, the work in [8] proposed a probit model with latent variables following a mixture model, which can successfully characterize the ordinality of data without the need to estimate these parameters.

In this paper, we propose informative priors on the parameters in the probit model, especially for the boundaries associated with category responses based on a Dirichlet distribution via a parameter transformation. The resulting posterior distributions with appropriate hyperparameter values chosen makes the sampling algorithm have fast convergence and efficient estimations. The rest of the article is arranged as follows. Section 2 presents our Bayesian inference procedure for the probit model with the latent variable approach. Subsequently, we carry out simulation studies to investigate the performance of the proposed methods in Section 3. For illustrative purposes, two real datasets are analyzed in Section 4, followed by some concluding remarks in Section 5.

2. Bayesian Method for the Probit Model

We first briefly introduce the probit model with a latent variable and then present a Bayesian procedure for the analysis. Let \((Z, X)\) denote the observed data, where the \(n\)-vector \(Z = (Z_1, Z_2, ..., Z_n)'\) is the \(n\)-ordered categorical outcomes and \(n \times p\) matrix of covariates \(X = (x_1, x_2, ..., x_n)'\) with \(x_i\) being the \(i\)th vector of covariates containing 1 in the first element. The responses \(Z_i\) take one of the \(J\) values, \(1, \ldots, J\), and are associated with the covariate vector \(x_i\) through a latent continuous variable \(y_i\) in the following linear regression model:

\[
y_i = x_i' \beta + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n,
\]

where \(\beta = (\beta_1, \beta_2, ..., \beta_p)'\) is a \(p \times 1\) vector of regression coefficients and \(\sigma^2\) is set to 1 to make the model identifiable. The correspondence between the observed outcome \(z_i\) and the latent variable \(y_i\) is defined by:

\[
z_i = j \text{ if } \delta_{j-1} < y_i \leq \delta_j, \quad j = 1, \ldots, J,
\]

where the boundaries \(\delta_j\) are unknown and \(-\infty = \delta_0 < \delta_1 < \ldots < \delta_{J-1} < \delta_J = \infty\).

2.1. Prior Specification

We chose an independent prior for the parameters \(\beta\) and \(\delta\), respectively. First, we specify a conjugate prior, multivariate normal, for the regression coefficients:

\[
\beta \sim N(\beta_0, \Sigma_0),
\]

...
with the $p$-vector mean $\beta_0$ and $p \times p$ variance-covariance matrix $\Sigma_0$ being hyperparameters. For setting a prior for the boundaries $\delta = (\delta_1, \delta_2, \ldots, \delta_{J-1})$, first, suppose that there is a continuous distribution function $F(\cdot)$ whose domain lies in $(-\infty, \infty)$, for example, a normal function, such that the coverage probability at each interval is $p_j = P(\delta_{j-1} < \Delta < \delta_j) = F(\delta_j) - F(\delta_{j-1}), j = 1, 2, \ldots, J$, and so, $\sum_{j=1}^{J} p_j = 1$. It follows that:

$$\begin{align*}
\begin{cases}
p_1 &= F(\delta_1) \\
p_2 &= F(\delta_2) - F(\delta_1) \\
\vdots \\
p_{J-1} &= F(\delta_{J-1}) - F(\delta_{J-2}) \\
p_J &= F(\delta_J) - F(\delta_{J-1})
\end{cases}
\iff
\begin{cases}
\delta_1 &= F^{-1}(p_1) \\
\delta_2 &= F^{-1}(p_1 + p_2) \\
\vdots \\
\delta_{J-1} &= F^{-1}(p_1 + p_2 + \cdots + p_{J-1}).
\end{cases}
\end{align*}$$

(6)

Secondly, a Dirichlet prior distribution is set for $(p_1, p_2, \ldots, p_J)$, that is, the prior density $\pi(p_1, p_2, \ldots, p_J | \gamma) = 1/B(\gamma) \times \prod_{j=1}^{J} p_j^{\gamma_j-1}$ with the positive hyperparameters $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_J)$ and the Beta coefficient $B(\gamma) = \prod_{j=1}^{J} \Gamma(\gamma_j)/\Gamma(\sum_{j=1}^{J} \gamma_j)$ with the gamma function $\Gamma(\cdot)$. Thus, by the transformation in (6), the prior of $\delta = (\delta_1, \delta_2, \ldots, \delta_{J-1})$ becomes:

$$\pi(\delta_1, \delta_2, \ldots, \delta_{J-1} | F, \gamma) = \frac{1}{B(\gamma)} \prod_{j=1}^{J} [F(\delta_j) - F(\delta_{j-1})]^{\gamma_j-1} \prod_{j=1}^{J-1} f(\delta_j),$$

(7)

where $f(\cdot)$ is the density function of $F(\cdot)$. Clearly, the expression (7) can be regarded as the joint density of order statistics $(\delta_1, \delta_2, \ldots, \delta_{J-1})$ with the rank index $(\gamma_1, \gamma_2, \ldots, \gamma_J)$ drawn from the distribution function $F(\cdot)$ (see, for example, [9]).

2.2. Posterior Inference

The prior beliefs are then updated with information from the data to lead to the following joint posterior distribution:

$$\pi(\beta, \delta | X, Y, Z) \propto L(\beta, \delta | X, Y, Z) \pi(\beta) \pi(\delta),$$

(8)

where $Y = (y_1, y_2, \ldots, y_n)'$ is the vector of latent variables and the likelihood function $L(\beta, \delta | X, Y, Z) = \prod_{i=1}^{n} f_Y(y_i | x_i, \beta) I(\delta_{j-1} < y_i < \delta_j, z_i = j)$ with $f_Y(\cdot)$ being the density function of $N(x_\beta, 1)$ and $I(\cdot)$ being the indicator function. Then, the conditional posteriors are:

$$\pi(\beta | X, Y, Z, \delta) \propto L(\beta, \delta | X, Y, Z) \pi(\beta),$$

resulting in $\beta | (X, Y, Z, \delta) \sim N(\tilde{\beta}, \tilde{\Sigma})$,

(9)

$$\pi(\delta | X, Y, Z, \beta) \propto \pi(\delta) \times \prod_{i=1}^{n} I(\delta_{j-1} < y_i < \delta_j, z_i = j),$$

(10)

with the posterior mean vector $\tilde{\beta} = (X'X + \Sigma_0^{-1})^{-1}(X'Y + \Sigma_0^{-1}\beta_0)$ and variance-covariance matrix $\tilde{\Sigma} = (X'X + \Sigma_0^{-1})^{-1}$. The conditional posterior of $\delta$ in (10) is a truncated joint density of order statistics $(\delta_1, \delta_2, \ldots, \delta_{J-1})$. To draw its posterior samples conveniently, we explore the conditional posterior distribution of each component $\delta_j$. Let $\delta_{(-j)}$ be the vector $\delta$ without the $j$th element, that is, $\delta_{(-j)} = (\delta_1, \ldots, \delta_{j-1}, \delta_{j+1}, \ldots, \delta_{J-1})$, and it follows that the conditional posterior of $\delta_j$ is:

$$\pi(\delta_j | X, Y, Z, \delta_{(-j)}) \propto [F(\delta_j) - F(\delta_{j-1})]^{\gamma_j-1} [F(\delta_{j+1}) - F(\delta_j)]^{\gamma_{j+1}-1} f(\delta_j) \times I(c_{j1} < \delta_j < c_{j2}), j = 1, 2, \ldots, J-1,$$

(11)

where $c_{j1} = \max\{y_i, i = 1, 2, \ldots, n : z_i = j\}, c_{j2} = \min\{y_i, i = 1, 2, \ldots, n : z_i = j+1\}$, $j = 1, 2, \ldots, J-1$, and $F(\delta_0) = 0, F(\delta_J) = 1$. Therefore, conditionally, $\delta_j$ is a random variable whose transformed $F(\delta_j)$ is distributed as the shifted $F(\delta_{j-1})$ and scaled $[F(\delta_{j+1}) - F(\delta_{j-1})]$ Beta distribution $\text{Beta}(\gamma_{j+1}, \gamma_{j+1})$ truncated at interval $[F(c_{j1}), F(c_{j2})]$, or equivalently:
\[
\frac{F(\delta_j) - F(\delta_{j-1})}{F(\delta_{j+1}) - F(\delta_j)} \mid (\delta_{j-1}, \delta_{j+1}) \sim \text{Beta}(\gamma_j, \gamma_{j+1})
\]

truncated at

\[
\begin{bmatrix}
F(c_{j,1}) - F(\delta_{j-1}) & F(c_{j,2}) - F(\delta_{j-1}) \\
F(\delta_{j+1}) - F(\delta_{j-1}) & F(\delta_{j+1}) - F(\delta_{j-1})
\end{bmatrix}.
\]

(12)

The method that we propose here is closely related to the approach presented in [10] for multinomial probit models. In this context, however, the correspondence between \(Z_i\) and \(Y_i\) uses different boundaries that account for the natural ordering of the outcome. We performed posterior inference using Markov chain Monte Carlo (MCMC) techniques. Specifically, a Gibbs sampling based on the above conditional posteriors was adopted starting from a set of initial values of parameters, and then, the following step was repeated \(M\) times, among which, given the values of \(Y^{(k)}, \beta^{(k)}\) and \(\delta^{(k)}\) at the \(k\)th iteration, the \((k+1)\)th iteration is as follows:

1. Update the latent vector \(Y\) from its posterior distribution given \((\beta, \delta, X, Z)\), each element of which is a truncated normal under the constraints defined in Equation (4):

\[
Y^{(k+1)}_i | (\beta^{(k)}, \delta^{(k)}, X, Z) \sim N(x_i^T \beta^{(k)}, 1), \quad \delta^{(k)}_{j-1} < Y^{(k+1)}_i \leq \delta^{(k)}_j \text{ if } Z_i = j, 
\]

\(j = 1, \ldots, J, i = 1, 2, \ldots, n.\)

(13)

2. Update the regression coefficients \(\beta\) from their posterior distribution in Equation (9) under the updated \(Y^{(k+1)}\).

3. Update the boundary parameters \(\delta_j\) from their posterior densities given in Equation (12) with \(c_{j,1}\) and \(c_{j,2}\) being evaluated at \(Y^{(k+1)}\). Specifically, first draw a \(d^{(k+1)}_j\) from the truncated Beta in Equation (12), and then, get \(\delta^{(k+1)}_j = F^{-1}(u^{(k+1)}_j)\) where \(u^{(k+1)}_j = F(\delta^{(k+1)}_{j-1}) + [F(\delta^{(k+1)}_j) - F(\delta^{(k+1)}_{j-1})] d^{(k+1)}_j, j = 1, 2, \ldots, J - 1.

2.3. Hyperparameter Settings

The prior on \(\beta\) depends on the mean \(\beta_0\) and covariance matrix \(\Sigma_0\), and the work in [11] discussed the relative merits and drawbacks of different specifications. Here, we set \(\beta_0 = 0\) and \(\Sigma_0 = cI\), which is easier to calibrate. The parameter \(c\) regulates the amount of shrinkage in the model. In general, we want to avoid very small values of \(c\), which cause too much regularization, and large values, which can induce nonlinear shrinkage as a result of Lindley’s paradox [12]. In the context of probit models for classification into nominal groups, the work in [10] provided some guidelines on how to choose this hyperparameter value. We used similar guidelines here. In practice, the values of \(c\) that provide good mixing of the MCMC sampler, with 25–50% distinct visited models, are appropriate [13]. An informative prior of \(\delta\) in (7) can be specified by using the number of categorical data to set the component \(\gamma_j\) in the parameter \(\gamma\) to be the counts in the \(j\)th category among \(z_i\)’s. Alternatively, a diffuse prior can be made by assigning all \(\gamma_j = 1\) to express no prior belief for \(\delta_j\), namely, a uniform distribution. For the transformed distribution function \(F(\cdot)\), we chose a normal distribution with zero mean and large-scale \(c_0\) (for example \(c_0 = 50\)) to cover a fairly wide range in \((-\infty, \infty)\), and so, in this case, \(F(\delta) = \Phi(\delta/c_0)\) with \(\Phi(\cdot)\) being the cumulative distribution function of the standard normal.

2.4. Posterior Prediction

The MCMC procedure results in a list of sampled \(Y, \beta,\) and \(\delta\) vectors. In order to draw posterior inference, we first need to impute the latent vector \(Y\), which can be viewed as missing data. Let \(\tilde{Y}, \tilde{\beta}\) and \(\tilde{\delta}\) be the estimates obtained by averaging respectively over the sampled \(Y, \beta,\) and \(\delta\) vectors.

Inference on class prediction can be done in various ways. If a future further vector of covariates \(x_f\) is available for validation, the least squares prediction based on a single value of \(\beta\) can be computed as:
\[ \hat{y}_f = x_f^T \hat{\beta} \quad \text{or} \quad \hat{y}_f = x_f^T \tilde{\beta}, \]  

where the posterior mean \( \tilde{\beta} = (X'X + \Sigma_0^{-1})^{-1}X'Y \). Alternatively, we can use the Bayesian model averaging over a set of a posteriori likely models of posterior samples \( \beta^{(k)} \) to estimate \( y_f \) as:

\[ \hat{y}_f = \sum_{k=1}^{M} x_f^T \beta^{(k)} \pi(\beta^{(k)} | X, \hat{Y}, Z, \delta). \]  

The ordered categorical outcomes can then be predicted using the boundary correspondence:

\[ \hat{z}_f = j \quad \text{if} \quad \delta_{j-1} < \hat{y}_f \leq \delta_j, \quad j = 1, 2, ..., J. \]  

In addition, by the fact that \( y_f \sim N(x_f^T \beta, 1) \), the prediction probability that it falls in each class can be computed through the model averaging over the posterior samples \( \beta^{(k)} \):

\[ P(Z_f = j) \approx \sum_{k=1}^{M} P(\delta_{j-1} < Y_f \leq \delta_j | \beta^{(k)}) \pi(\beta^{(k)} | X, \hat{Y}, Z, \delta) \]

\[ = \left[ \Phi(\delta_j - x_f^T \beta^{(k)}) - \Phi(\delta_{j-1} - x_f^T \beta^{(k)}) \right] \pi(\beta^{(k)} | X, \hat{Y}, Z, \delta), \]  

where \( \Phi(\cdot) \) is the distribution function of the standard normal distribution. The class membership can then be predicted by the mode of the predictive distribution:

\[ \hat{z}_f = \arg\max_{1 \leq j \leq J} P(Z_f = j). \]  

Furthermore, a less varied estimate can be obtained by the nearest rounded integer computed through membership averaging over predictive probabilities,

\[ \hat{z}_f = \lfloor \mu_Z \rfloor, \quad \text{where} \quad \mu_Z = \frac{1}{J} \sum_{j=1}^{J} jP(Z_f = j). \]  

### 3. Simulation Study

We conducted a simulation study to assess the performance of the proposed Bayesian method. The simulated data sets were from three multivariate normal distributions whose dimension was four and the means \( \mu_1 = [3, 2, 4, 1]' \), \( \mu_2 = [3, -2, 4, -1]' \), \( \mu_3 = [-3, -2, -4, -1]' \) and equal correlation variance-covariance matrices \( \Sigma = \sigma^2[(1 - \rho)I + \rho 11'] \), where \( \sigma = 2 \), the increasingly-ordered correlations \( \rho = 0.1, 0.5, 0.9 \) for three structures, corresponding to the ordered response \( z = 1, 2, 3 \), respectively, \( I \) is the identity matrix, and \( 1 \) is the vector of ones. We simulated 30 observational data \( x = (x_1, x_2, x_3, x_4)' \) from each multivariate normal to make a total sample size of 90. Considering the structure of the data may be an essential aspect for efficient classification. In order not to make erroneous conclusions from the predicted error rates, we critically examined the nature of the data to explore the variations existing among groups. In a model for predicting ordinal outcomes, it is obvious to believe that, remarkably, the least variations and far apart locations among groups of variables could make it easy to classify or predict correctly into which category a particular observation falls. In this vein, we visualized the data structure by using the tool of displaying data concentration in [14], who proposed the \( p \)-dimensional ellipsoid, \( E_d \), of size (“radius”) \( d \), which is defined as the set of all points \( X \) in a contour:

\[ E_d(X, S) = \{ X : (X - X)^T S^{-1}(X - X) \leq d^2 \}. \]
Clearly, \( E_d \) corresponds to the set of points whose Mahalanobis distances \( D^2 = (X - \bar{X})' S^{-1} (X - \bar{X}) \) with sample covariance matrix \( S \), from the centroid of the sample, \( \bar{X} = (\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_p)' \), are less than or equal to \( d^2 \). For multivariate normal variables, the data ellipsoid approximates a contour with constant density in their joint distribution, and \( D^2 \) is the asymptotic chi-squared distribution with \( p \) degrees of freedom, \( \chi^2_p \). The work in [15] elaborated more on the properties of data ellipsoids and their use in a wide variety of problems and applications. Taking \( d^2 = \chi^2_{0.05, 2} = 5.99 \), the two-dimensional pairwise ellipsoid is shown in Figure 1, where the ellipse encloses approximately 95% of the data points under the normal theory. The plot indicates that there were some overlaps between Groups 1 and 2, and Group 3 was relatively well-separated away from the other two. Thus, we expected a slightly high classification error rate where the most misclassified cases would probably occur among Groups 1 and 2.

Next, we performed the analysis by our method. To obtain accurate results and avoid over-fitting, we applied the Bayesian model on the simulated data (training data), and generated a test data with 300 observations for each category for validation. We ran four MCMC chains with widely different starting values for 10,000 iterations each and discarded the first 2000 as burn-in to eliminate dependence on the starting points. We also considered several hyperparameter values for the covariance of the regression coefficients \( \Sigma_0 = cI \), with \( c \) ranging between 5 and 20. There was minimal effect on the overall results, and here, we report the results for \( c = 10 \). An informative prior for the boundary parameter vector \( \gamma \) was specified by setting all the components \( \gamma_j = 30 \), the counts in the \( j \)th category among ordinal responses. We note that despite the widely different starting values, there was a good agreement between the results. The error rate for the test data was recorded from the predictions to the misclassification rate tabulated in Table 1, containing the three prediction approaches in Equations (16) (where \( y_j \) is estimated in Equations (14)), (18) and (19). The classification result showed that there were in total about 270 subjects misclassified, among which, 90 out of 300 observations in Group 1 were misclassified into Group 2, while 100 obs in Group 2 were incorrectly assigned to Group 1. The outcome validated our previous judgment from the pairwise ellipses in Figure 1, where Groups 1 and 2 shared quite a bit of common area. As shown in Table 1, the three types of predictions approximately yielded the same error rates, and the polytomous ordinal logistic regression (POLR) model resulted in a higher classification error rate.

Finally, for comparison purposes, we analyzed the data using common classification methods, such as linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), \( k \)-nearest neighbor (KNN), and support vector machines (SVM), which build multi-class classifiers without taking the
natural ordering of the response into account. For KNN, we considered values of group number \( k \) ranging from 2–8, and we report the results for \( k = 3 \), which gave the lowest overall misclassification rate. All these approaches produced a higher error rate as summarized in Table 1.

Table 1. Simulated data: test data prediction misclassification rates. POLR, polytomous ordinal logistic regression; QDA, quadratic discriminant analysis.

<table>
<thead>
<tr>
<th>Bayesian Method</th>
<th>POLR</th>
<th>LDA</th>
<th>QDA</th>
<th>KNN</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boundary Probability Average</td>
<td>0.308</td>
<td>0.307</td>
<td>0.294</td>
<td>0.348</td>
<td>0.362</td>
</tr>
</tbody>
</table>

4. Real Data Analysis

We applied our model for discrimination analysis on two real datasets. The first was the well-known Irisflower data, and the second was about the measurements of male Egyptian skull data. The purpose of adopting these two datasets was to examine how the model predictions were affected by different data structures and measurement variations among groups.

4.1. Iris Data

We used the dataset as a benchmark for the analysis. It was introduced by the British statistician Ronald [16]. It is also sometimes called Anderson’s Iris data [17], which was collected to quantify the morphological variation of Iris flowers of three related species pictured in Figure 2. The dataset consisted of 50 samples from each of three species (Iris setosa, Iris versicolor, and Iris virginica) with four measured features for each sample: the length and width of the sepals and petals, in centimeters. These are generally viewed as nominal category data and have been analyzed in a vast literature, such as [18,19], and many others. Here, we explore it as ordinal outcomes of three species instead, ranked by the magnitude of measurement variations. The total variances of the four morphological measurements for each species were: 0.3292 (setosa), 0.6248 (versicolor), and 0.8883 (virginica). These numbers, to some extent, represent the size and spread of flowers, which is consistent with the image in Figure 2, where the flower of setosa has the least size in terms of sepal and petal, versicolor larger, and virginica largest.

These findings can also be observed from the visual display of the ellipses for all pairs of measurements shown in Figure 3, where one may further notice that, except a little overlap between groups of versicolor and virginica, the three species of data were overall well-separated; especially, Iris setosa was far apart from the other two species. To obtain an accurate error rate of classification, we applied the cross-validation approach to partition the whole dataset into a training set containing 120 observations with 40 for each category and a testing data containing 30 observations with 10 for each category. The partitioning was repeated five times until all the samples from each species were exhausted. The small error rate of classification produced by our method with cross-validation prediction, along with POLR regression, is summarized in Table 2, where all the cases of setosa were classified correctly by both methods, two (or three) cases from versicolor were misclassified to the virginica species by the Bayesian (or POLR) method, and one from virginica was incorrectly assigned to versicolor by the POLR model. Other common classification methods dealing with nominal data yielded a little larger error rate, shown in Table 2. To see how much the reduction of the error rates for each prediction approach were, we simply computed the error rate of the null model where no covariates was used, that is, \( 1 - \max(\text{RF}_j) \), as listed in Table 2, where \( \text{RF}_j \) is the relative frequency for the \( j \)th category in the dataset.
**Figure 2.** Three species of Iris flower.

**Figure 3.** Iris data: ellipse structure of three species.

**Table 2.** Real data: cross-validated prediction misclassification rates.

<table>
<thead>
<tr>
<th>Bayesian Method</th>
<th>POLR</th>
<th>LDA</th>
<th>QDA</th>
<th>KNN</th>
<th>SVM</th>
<th>Null Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris</td>
<td>0.013</td>
<td>0.013</td>
<td>0.013</td>
<td>0.026</td>
<td>0.036</td>
<td>0.034</td>
</tr>
<tr>
<td>Skull</td>
<td>0.553</td>
<td>0.546</td>
<td>0.533</td>
<td>0.607</td>
<td>0.646</td>
<td>0.633</td>
</tr>
</tbody>
</table>
4.2. Skull Data

The dataset was obtained from the R embedded package “HSAUR”. The data consisted of four physical measurements in millimeters of 30 male Egyptian skulls from each of five epochs (periods) [20]: Period 1 (4000 BC), Period 2 (3300 BC), Period 3 (1850 BC), Period 4 (200 BC), and Period 5 (150 AD). The measures were the maximal breadth (mb), basibregmatic height (bh), basialveolar length (bl), and nasal height (nh) of each skull. Figure 4 gives an illustrative labeled image for these four measurements of a typical skull. The researchers claimed that a change in skull measurement was as a result of the time duration. Systematic changes over time could indicate interbreeding among migrant populations (or the influence of other factors) [21]. The interest in this analysis, however, lies in the ability to predict which period these measurements fall within well.

Figure 4. A labeled male Egyptian skull.

Figure 5 displays the ellipses for all pairs of measurements, and it clearly shows that the five groups of data were pretty much overlapping each other so that one can hardly distinguish one class from the others, and a high classification error rate can be expected. For our Bayesian method, an MCMC chain with 10,000 iterations was run, and the first half was discarded to eliminate dependence on the starting points. Several hyperparameter values were chosen for the covariance of the regression coefficients, with $c$ ranging between 5 and 15. We note that different hyperparameter values resulted in a similar classification. The error rates of cross-validated prediction are listed in Table 2, where, not surprisingly, the high error rates (83, 82, and 80 skulls were misclassified respectively by the three prediction approaches of the Bayesian method, while 92 skulls were incorrectly assigned by POLR) of classification produced by our method validated our previous judgment. In contrast to the Iris data, the skull dataset provided an example where the measurements tended to cluster around a common centroid and overlapped one another with little evidence of separation in location. The classification results of these two different structural datasets justified the reliability of our model in exploration and prediction for ordinal outcomes’ data.

We also noticed that all other common procedures even yielded higher error rates compared to our method summarized in Table 2. As a result of all these classification outcomes, we could logically infer that the measurements of these male Egyptian skulls did not change significantly across the given time periods.
5. Concluding Remarks

We have proposed a Bayesian approach for predicting ordinal outcomes. Introducing latent variables underlying the ordinal outcomes, the problem reduces to a linear model setting. While MCMC techniques are generally computationally intensive, with an informative prior posed on the boundary parameters, our MCMC algorithm showed practical implementation and fast convergence. The simulation study demonstrated the efficient and impressive performance of the proposed method. We have illustrated that our approach, with applications to two real datasets, provided an efficient, reliable, and precise analysis for ordinal categorical response data.

Author Contributions: N.S. contributions: methodology proposing and development; theoretical derivation; paper organization and writing; B.O.D. contribution: code writing and running; paper writing.

Funding: This research was funded by NSF CMMI-0654417 and NIMHD-2G12MD007592.

Conflicts of Interest: The authors declare no conflict of interest.

References

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).