

DIRICHLET PROCESS MIXED MODEL FOR BIVARIATE
ORDERED CATEGORICAL DATA WITH APPLICATION TO THE
WISCONSIN EPIDEMIOLOGIC STUDY OF DIABETIC
RETINOPATHY

Kalyan Das¹ and Atanu Biswas²

¹Department of Statistics, University of Calcutta
35 Ballygunge Circular Road, Calcutta - 700 019, India

e-mail : kalyan@cubmb.ernet.in

²Applied Statistics Unit, Indian Statistical Institute
203 B. T. Road, Calcutta - 700 035, India

e-mail : atanu@isical.ac.in

Abstract: The present article proposes a latent variable based mixed model for bivariate ordered categorical data in a Bayesian framework. The proposed model allows for random effects and covariates for the margins. A Dirichlet process mixing through the random effects adds considerable flexibility to this model. Recent computational advances enable them to be fitted easily. Illustration of the proposed model uses data from Wisconsin Epidemiologic Study of Diabetic Retinopathy for identifying risk factors for diabetic retinopathy among younger-onset diabetics. Some other examples are given where the proposed model is applicable. An analysis has also been indicated in situations where a few cells in the bivariate categorical data are not available.

Some Key Words and Phrases: Bivariate ordinal data, Latent variable, Gibbs sampler, Dirichlet process prior, Monte Carlo Markov Chain.

AMS 1991 subject classification: Primary 62J12; Secondary 62F15.

1. Introduction

In many areas of medical and social sciences, two-way contingency tables for paired data with natural ordering in both margins, occur frequently. The responses in each pair are recorded on an ordinal scale, for example, mild, moderate, severe etc. (see for example, Ashford (1959), Cox (1970), McCullagh (1980) and Snell (1964)). The ordered responses in each pair may be clustered and hence the responses of the individuals within the clusters can be positively correlated. Further there are many covariates measured not only for the pair but also for each member of the pair separately. Our immediate concern is to develop a flexible model which describes the relationship between bivariate ordered categorical responses and the various covariates available.

Eversince Dale (1986) proposed the analysis of bivariate ordinal categorical data, a considerable studies have been made in this fascinating research area in statistics. Molenberghs and Lesaffre (1994) have used a multivariate Plackett distribution for the extension of Dale model. Recently, Williamson, Kim and Lipsitz (1995) have used the generalized estimating equations approach as an alternative to computationally expensive likelihood methods considered by Dale (1986) and Molenberghs and Lesaffre (1994). (See also Kim, 1995, Lesaffre and Molenberghs, 1991, in this context.) But most of the models are not as flexible as desired and in fact none of these models has taken into account the severity or the cluster effects.

We have considered here a semiparametric Bayesian approach for the analysis of latent variable based bivariate ordinal categorical model that allows random effects and subject specific (continuous, binary or count) covariates. Some examples have been

chosen to illustrate the wide range of applications of the model. The first example is an ophthalmologic study where the bivariate ordinal responses are observed. The second example is a study on anxiety scores (overt and covert) which are also recorded in an ordered categorical manner. The third example is a study relating to depth of a river bed where the pre- and post-monsoon depth are actually the ordered categorical responses.

In the Bayesian paradigm a nonparametric distribution of the random effects has been considered earlier by West et al. (1994) in the context of repeated measures and Bush and MacEachern (1996) in the context of randomized complete block designs. We provide a general framework for Bayesian analysis of mixed effects models in which a nonparametric Dirichlet process prior for the random effects has been taken into consideration. This may better express the uncertainty about the true distribution of the random effects.

Essentially in the nonparametric Bayesian approach, the usual normal prior for the random severity factor b_i has been replaced by a nonparametric distribution followed by a Dirichlet process prior as the general prior distribution for b_i . The fundamental foundational work has been given in Ferguson (1973). Its application, particularly in Gibbs sampler, has been done by Doss (1994), MacEachern (1994), Escobar (1994), Bush and MacEachern (1996), Liu (1996), Muller et al. (1996) and MacEachern and Muller (1998).

The examples mentioned above are discussed in details in section 2. In section 3, Dirichlet process mixed models for bivariate categorical data are presented and the method of analyzing such data in a Bayesian framework is described. An ophthalmologic dataset from the Wisconsin Epidemiologic Study of Diabetic retinopathy (WESDR) are analyzed in section 4. The advantages of our method over the classical probit analysis of the data performed by Kim (1995) are also discussed there. Finally,

in section 5, a more general set up is considered where a few cells in the two way contingency table ceases to be present. Some concluding remarks are made at the end.

2. Examples

In several biomedical studies categorical outcomes are quite common. Ordinal scales for measurement are often used in the absence of well defined non-invasive direct measurements. One classical example of bivariate categorical responses arises in the context of ophthalmologic studies, where measurements on both eyes are taken in an ordered categorical fashion. For example, one often considers retinopathy of the two eyes which are generally measured in four ordered categories namely no retinopathy, mild nonproliferative retinopathy, moderate to severe nonproliferative retinopathy and proliferative retinopathy, respectively. Several person- and eye-specific covariates also dictate the retinopathy levels. The Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR) dataset on 996 insulin-taking younger-onset diabetics, reported by Klein et al. (1984), has earlier been analyzed in various ways by some leading statisticians.

In some psychological studies anxiety scores are of interest. The anxiety levels are scored with the help of the Anxiety Scale Questionnaire (ASQ) (see Krugg et al. (1976)) that consists of 40 questions divided into two parts. Consequently, the total anxiety score (based on responses to all the 40 questions) has a breakdown into *overt*, i.e., symptomatic anxiety score (based on responses to 20 questions) and a *covert*, i.e., unrealized anxiety scores (based on responses to 20 questions). Often the overt and covert scores are further categorized in an ordinal scale for better understanding, and we are left with a similar bivariate ordinal set up. Several person-specific covariates

influence such anxiety scores. See Mukhopadhyay (1989) for details.

In some geographical studies interest may be to look into the depth of river beds at different points of time. Siltation is a continuous natural process by which the depth and water carrying capacity of a river bed decreases through accumulation of silt, sand, pebbles etc. in the river-bed. This natural process is mainly due to a number of (identifiable) random causes like rainfall, amount of discharge of water through river-bed, geological characteristics of soil of the catchment area and path of the river etc. Both pre- and post-monsoon depths of the river bed may be of interest in several studies, and these depths are often observed in an ordered categorical manner. An experiment was conducted by the Calcutta Port Trust authority during 1990-1994 on the part of river the Ganges in between two places Cossipore and Gardenrich in Calcutta.

There may be situations when due to some reason, observation in some cells may not be available. Classical likelihood based analysis of such a truncated bivariate ordinal data seems to be extremely difficult (see Weiss, 1993). A brief discussion on this kind of data has been made in section 5.

3. Bivariate Ordinal Model and its Analysis

3.1. The model

The concept of latent variables is often used for binary or categorical responses. Snell (1964) postulated the idea of continuous latent variables in the context of linear logistic models for ordered categorical data. several authors have used this idea in the

analysis of their problems. McCullagh (1980) and Muthen (1984) and others have used them in the context of ordinal response model. Let y be a categorical response and y^* be an underlying latent variable. Then one observes only the categorical variable y with

$$y = k \quad \text{if} \quad \gamma_{k-1} < y^* \leq \gamma_k,$$

where the cut-off points γ_{k-1} , γ_k are usually unknown. In the context of analyzing ophthalmologic data of the WESDR study, Kim (1995) extended the latent variable technique to a bivariate probit regression model. Also Weiss (1993) used a truncated bivariate ordinal model in his study of a dataset on motorcycle injuries in Los Angeles. The latent variable technique for the bivariate case can be considered as follows.

Let y_{1i} and y_{2i} denote the bivariate ordered categorical responses on two components for the i -th subject. Suppose y_{1i} can take values $0, 1, 2, \dots, k_1$ and similarly y_{2i} can take values $0, 1, 2, \dots, k_2$. Corresponding to each pair (y_{1i}, y_{2i}) , we can think of a pair of latent variables (y_{1i}^*, y_{2i}^*) such that

$$\begin{aligned} y_{1i} = j & \quad \text{if} \quad \gamma_{1j} < y_{1i}^* \leq \gamma_{1j+1}, \quad j = 0, 1, \dots, k_1, \\ y_{2i} = l & \quad \text{if} \quad \gamma_{2l} < y_{2i}^* \leq \gamma_{2l+1}, \quad l = 0, 1, \dots, k_2, \end{aligned}$$

where the two sets of cut-off (break) points $\gamma_1 = (\gamma_{10}, \dots, \gamma_{1k_1+1})$ and $\gamma_2 = (\gamma_{20}, \dots, \gamma_{2k_2+1})$ are unknown. To avoid complications, we take $\gamma_{10} = \gamma_{20} = -\infty$ and $\gamma_{1k_1+1} = \gamma_{2k_2+1} = \infty$. Thus the pair of responses (y_{1i}, y_{2i}) can take $(k_1 + 1) \times (k_2 + 1)$ possible values (j, l) where $(j, l) \in S_1 = \{(j, l) : j = 0, 1, \dots, k_1, l = 0, 1, \dots, k_2\}$. We assume that (y_{1i}^*, y_{2i}^*) follow a bivariate model

$$y_{1i}^* = x'_{1i}\beta_1 + z'_{1i}b_{1i} + \epsilon_{1i},$$

$$y_{2i}^* = x_{2i}'\beta_2 + z_{2i}'b_{2i} + \epsilon_{2i}, \quad (3.1)$$

where x_{1i} and x_{2i} are the vector of covariates for the i -th subject for the two components of the bivariate response, the disturbance vector $\epsilon_i = (\epsilon_{1i}, \epsilon_{2i})'$, $i = 1, \dots, n$, are independently distributed as $N_2(0, \Sigma)$, $\Sigma = ((\sigma_{ss'}))$, $s, s' = 1, 2$. Also b_{1i} and b_{2i} are vectors of random effects (or the random severity factors) with associated design vectors being z_{1i} and z_{2i} respectively.

Denote

$$\beta = (\beta_1', \beta_2')', \quad \gamma = (\gamma_1', \gamma_2')', \quad b = (b_1', b_2')',$$

the response vector

$$y = (y_1', \dots, y_i', \dots, y_n')', \quad y_i = (y_{i1}, y_{2i})',$$

and similarly the latent vector

$$y^* = (y_1^{*'}, \dots, y_i^{*'}, \dots, y_n^{*'})', \quad y_i^* = (y_{i1}^*, y_{2i}^*)'.$$

Clearly, y_i^* can be expressed as

$$y_i^* = X_i\beta + Z_i b_i + \epsilon_i,$$

where

$$X_i = \begin{pmatrix} x_{1i}' & 0 \\ 0 & x_{2i}' \end{pmatrix} \text{ and } Z_i = \begin{pmatrix} z_{1i}' & 0 \\ 0 & z_{2i}' \end{pmatrix}.$$

Thus there are $p (= p_1 + p_2)$ fixed effects and $r (= r_1 + r_2)$ random effects in the model.

Then, conditional on β , Σ , γ and b , the joint distribution of y and y^* is

$$\pi(y^*, y | \beta, \Sigma, \gamma, b) = \prod_{i=1}^n \left[\sum_{j, l \in S_1} 1_{jl}^i 1(\gamma_{1j} < y_{1i}^* \leq \gamma_{1j+1}, \gamma_{2l} < y_{2i}^* \leq \gamma_{2l+1}) \right] \times N_2(X_i \beta + Z_i b_i, \Sigma), \quad (3.2)$$

where the set S_1 is defined earlier, $1(X \in A) = 1$ or 0 according as $X \in A$ or not and $1_{jl}^i = 1$ or 0 according as $y_{1i} = j, y_{2i} = l$, or not.

3.2: A Bayesian Framework

The prior belief of the experimenter can be set in the form of a suitable prior $\pi(\beta, \Sigma, \gamma)$ for β, Σ and γ . Observe that the conditional joint posterior density is therefore

$$\pi(\beta, \Sigma, \gamma, b | y^*, y) = c \pi(\beta, \Sigma, \gamma, b) \pi(b) \pi(y^*, y | \beta, \Sigma, \gamma, b), \quad (3.3)$$

where c is a generic constant and $\pi(b)$ is the marginal density of b . Note that, although it is difficult to work with (3.3), analytically it is possible to use Monte Carlo Markov Chain (MCMC) sampling to explore the posterior distributions of β, Σ, γ and b .

3.3: Markov Chain Monte Carlo Technique

To apply such a technique we look at the full conditionals. We assume a diffuse prior

$$\pi(\beta, \Sigma, \gamma) \propto |\Sigma|^{-3/2}, \quad (3.4)$$

and a Dirichlet Process (DP) prior is assumed for the unknown distribution G of b_i such that

$$\begin{aligned} b_i &\sim G, \\ G &\sim DP(\alpha, G_0), \end{aligned}$$

with

$$G_0 \sim N_2(0, \Gamma),$$

where Γ is a known dispersion matrix of order 2. When G_0 is known, i.e., G is a fully parametric prior, the posterior distributions can be easily obtained as in Wilks et al. (1993). Suppose the density of G_0 is g_0 . Analytic derivation of the posterior p.d.f. of β is quite hard and we do not proceed to do that. In view of (3.4), the conditional posterior of β can be obtained as

$$[\beta | \Sigma, \gamma, b, y^*, y] \sim N_p \left(\hat{\beta}_{y^*}, (X' \Omega^{-1} X)^{-1} \right), \quad (3.5)$$

where

$$\begin{aligned} \hat{\beta}_{y^*} &= (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} \tilde{y}^*, \\ \tilde{y}_i^* &= y_i^* - Z_i b_i, \quad \tilde{y}^* = (\tilde{y}_1^*, \dots, \tilde{y}_n^*)', \\ \Omega &= I_n \otimes \Sigma, \\ X &= (X_1 \ X_2 \ \dots \ X_i \ \dots \ X_n)', \end{aligned}$$

with X_i being a $2 \times p$ matrix. Also, the conditional posterior of Σ can be computed as

$$[\Sigma | \beta, \gamma, b, y^*, y] \propto |A|^{n/2} |\Sigma|^{-\frac{n+3}{2}} e^{-\frac{1}{2} \text{tr } \Sigma^{-1} A}. \quad (3.6)$$

Clearly, this is an inverted Wishart distribution with

$$A = \sum_{i=1}^n (y_i^* - X_i\beta - Z_i b_i) (y_i^* - X_i\beta - Z_i b_i)',$$

and finally the conditional posterior of $(\gamma_{1j}, \gamma_{2l})'$ is

$$\begin{aligned} \left[\begin{pmatrix} \gamma_{1j} \\ \gamma_{2l} \end{pmatrix} \middle| \beta, \Sigma, \gamma_{-(j)}', b, y^*, y \right] &\propto \prod_{i=1}^n \left[1_{j-1, l-1}^i \mathbb{1} \left(\begin{matrix} \gamma_{1j-1} < y_{1i}^* \leq \gamma_{1j} \\ \gamma_{2l-1} < y_{2i}^* \leq \gamma_{2l} \end{matrix} \right) + 1_{j, l-1}^i \mathbb{1} \left(\begin{matrix} \gamma_{1j} < y_{1i}^* \leq \gamma_{1j+1} \\ \gamma_{2l-1} < y_{2i}^* \leq \gamma_{2l} \end{matrix} \right) \right. \\ &\quad \left. + 1_{j-1, l}^i \mathbb{1} \left(\begin{matrix} \gamma_{1j-1} < y_{1i}^* \leq \gamma_{1j} \\ \gamma_{2l} < y_{2i}^* \leq \gamma_{2l+1} \end{matrix} \right) + 1_{j, l}^i \mathbb{1} \left(\begin{matrix} \gamma_{1j} < y_{1i}^* \leq \gamma_{1j+1} \\ \gamma_{2l} < y_{2i}^* \leq \gamma_{2l+1} \end{matrix} \right) \right], \end{aligned} \quad (3.7)$$

where $\gamma_{-(j)}$ is the collection of all γ_{1u}, γ_{2v} 's except γ_{1j} and γ_{2l} . This conditional distribution can be seen to be uniform over the region specified by the four points $a_1 = (d_1, d_2)'$, $a_2 = (d_3, d_2)'$, $a_3 = (d_1, d_4)'$, $a_4 = (d_3, d_4)'$ in the two dimensional plane, where

$$\begin{aligned} d_1 &= \max \{ \max \{ y_{1i}^* : y_{1i} = j - 1 \}, \gamma_{1j-1} \}, \\ d_2 &= \max \{ \max \{ y_{2i}^* : y_{2i} = l - 1 \}, \gamma_{2l-1} \}, \\ d_3 &= \min \{ \min \{ y_{1i}^* : y_{1i} = j + 1 \}, \gamma_{1j+1} \}, \\ d_4 &= \min \{ \min \{ y_{2i}^* : y_{2i} = l + 1 \}, \gamma_{2l+1} \}. \end{aligned}$$

For each i , the conditional distribution of y_i^* is

$$\left[y_i^* \middle| \beta, \Sigma, \gamma, b, y_i = \begin{pmatrix} j \\ l \end{pmatrix} \right] = N_2 (X_i\beta + Z_i b_i, \Sigma) \cdot \mathbb{1} \left(\begin{matrix} \gamma_{1j} < y_{1i}^* \leq \gamma_{1j+1} \\ \gamma_{2l} < y_{2i}^* \leq \gamma_{2l+1} \end{matrix} \right), \quad (3.8)$$

which is the truncated bivariate density with a rectangular domain specified by the four points $(\gamma_{1j}, \gamma_{2l})'$, $(\gamma_{1, j+1}, \gamma_{2l})'$, $(\gamma_{1, j+1}, \gamma_{2, l+1})'$ and $(\gamma_{1j}, \gamma_{2, l+1})'$ in the two dimensional plane.

We denote by b_{-i} the random effects of the observation vectors excluding that for the i -th observation. Then arguing as in Escobar (1994) and West et al. (1994), we write

$$[b_i | \beta, \Sigma, \gamma, b_{-i}, y^*, y] \propto \sum_{j \neq i} q_j \delta_{b_j} + \alpha q_0 g_0(b_i) p(y_i, y_i^* | b_i, \beta, \Sigma, \gamma), \quad (3.9)$$

where

$$\begin{aligned} q_0 &= \int p(y_i, y_i^* | b_i, \beta, \Sigma, \gamma) g_0(b_i | \Gamma) db_i, \\ q_j &= p(y_i, y_i^* | b_j, \beta, \Sigma, \gamma). \end{aligned}$$

Note that

$$p(y_i, y_i^* | b_i, \beta, \Sigma, \gamma) = p(y_i | y_i^*, b_i, \beta, \Sigma, \gamma) \times p(y_i^* | b_i, \beta, \Sigma, \gamma).$$

After some routine algebra we obtain

$$\begin{aligned} & [b_i | \beta, \Sigma, \gamma, b, y^*, y] \\ & \propto \left(\sum_{j \neq i} \exp \left\{ -\frac{1}{2} (y_j^* - X_i \beta - Z_j b_i)' \Sigma^{-1} (y_j^* - X_i \beta - Z_j b_i) \right\} \cdot \delta_{b_j} \right) \\ & + \alpha (2\pi)^{-r/2} |\Gamma|^{-1/2} |\Psi_i|^{-1/2} \exp \left\{ -\frac{1}{2} (y_i^* - X_i \beta)' [\Sigma^{-1} - \Sigma^{-1} Z_i \Psi_i Z_i' \Sigma^{-1}] (y_i^* - X_i \beta) \right\} \\ & \times g_0(b_i | \Gamma) p(y_i, y_i^* | b_i, \beta, \Sigma, \gamma), \end{aligned}$$

where

$$\Psi_i = (\Gamma^{-1} + Z_i \Sigma^{-1} Z_i')^{-1}.$$

Thus with probability proportional to

$$\exp \left\{ -\frac{1}{2} (y_i^* - X_i \beta - Z_i b_i)' \Sigma^{-1} (y_i^* - X_i \beta - Z_i b_i) \right\}$$

we select from distribution δ_{b_j} , which means that we select $b_i = b_j$. Also, with probability proportional to

$$\alpha(2\pi)^{-r/2}|\Gamma|^{-1/2}|\Psi_i|^{-1/2} \exp \left\{ -\frac{1}{2}(y_i^* - X_i\beta)' \left[\Sigma^{-1} - \Sigma^{-1}Z_i\Psi_iZ_i'\Sigma^{-1} \right] (y_i^* - X_i\beta) \right\}$$

we select from

$$p(b_i|\beta, \Sigma, \gamma, y_i, y_i^*) \propto g_0(b_i|\Gamma)p(y_i, y_i^*|b_i, \beta, \Sigma, \gamma).$$

This means we sample b_i from its full conditional,

$$[b_i|\beta, \Sigma, \gamma, b_{-i}, y^*, y] \sim N_r \left(\Psi_i Z_i' \Sigma^{-1} (y_i^* - X_i \beta), \Psi_i \right).$$

This results in a mixture distribution where one piece is a normal distribution and all other are point masses.

Now to implement the well known Gibbs sampler (Geman and Geman, 1984, Gelfand et al., 1990) we start with an initial guess at β, Σ, γ (may be MLE), y^* , b and simulate from the conditional distributions (3.5) – (3.9). Thus after a large number (t) of iterations we obtain a sample from $[\beta, \Sigma, \gamma, y^*, b|y]$, and work with such samples.

3.4. Prior for Γ :

In case Γ is unknown, we can assume some prior for Γ . Suppose

$$\Gamma^{-1} \sim \text{Wishart}(d, D_0),$$

$d \geq 2$, and D_0 is a 2×2 known positive definite matrix. Then

$$p(\Gamma^{-1}|d, G_0) \propto |\Gamma^{-1}|^{\frac{d-3}{2}} \exp \left\{ -\frac{1}{2} \text{trace} (D_0^{-1} \Gamma^{-1}) \right\}.$$

After choosing random effects for each subject, the subjects will be grouped into clusters in which the subjects have a common b_i . Suppose in the sample the number of distinct b_i 's are ν , and the distinct values are $\eta_1, \eta_2, \dots, \eta_\nu$. Then the conditional posterior distribution of all the other remains as usual. In addition, the posterior distribution of Γ^{-1} is obtained as

$$[\Gamma^{-1} | \beta, \Sigma, \gamma, b, y^*, y] \sim \text{Wishart} \left(d + \nu, \left(D_0^{-1} + \sum_{j=1}^{\nu} \eta_j \eta_j' \right)^{-1} \right).$$

4. WESDR Data and Computational Studies

In this section, we discuss elaborately the WESDR dataset and carry out a semi-parametric analysis.

The study was conducted to assess risk factors for diabetic retinopathy among younger-onset diabetics. Ocular examination was done for 996 insulin-taking younger-onset diabetics and the severity of diabetic retinopathy of those diabetics was graded according to a 10 point ordinal scale. These are then grouped into four categories e.g. no retinopathy, mild nonproliferative retinopathy, moderate to severe nonproliferative retinopathy, and proliferative retinopathy, separately for both the eyes. Thus we are left with a 4×4 squared arrangement of the data. Of the 996 samples a few covariates are missing from quite a large proportion of data. Accordingly we discard those samples from our analysis, and carry out our analysis based on 691 observations for which complete information are available.

In all, 3 eye-specific covariates are recorded along with 11 person-specific covari-

ates. The eye-specific covariates are the right and left eye macular edema (ME) (present/absent), right and left eye refractive error (RE) in diopters, right and left eye intraocular pressure (IOP) in mmHg., while the person-specific covariates are age at diagnosis (AgD) of diabetes in years, duration of diabetes (DuD) in years, glycosylated hemoglobin (GH) in percent, systolic and diastolic blood pressures (SBP & DBP) in mmHg., body mass index (BMI) in kilograms per meter squared, pulse rate (PR) in beats per 30 seconds, sex, urine protein (UP) (present/absent), doses of insulin (DI) per day and area of residence (AR) (urban/rural).

For computation we use the Gibbs sampler approach using WINBUGS (see <http://www.mrc-bsu.cam.ac.uk/bugs/> for details). A 4000 update burn in followed by a further 4000 updates provides the posterior summary statistics like mean, s.d., median and 95% probability interval. The Monte Carlo error involved in the computation is also reported in Table 1 for known Γ (= identity matrix). Some computations are also carried out using a prior for Γ with D_0 as the identity matrix. Here, for all the computations, we have considered $r_1 = r_2 = 1$, and $Z_i = I_2$.

The WESDR data has earlier been analyzed by Kim (1995) from a frequentist's point of view. Although in the present paper, we consider a similar bivariate probit function as in Kim (1995), our approach is somewhat different. Besides the basic philosophical difference, there is some additional merit in our model. In our model a random severity component has been introduced to explain the variability. Analysis based on Kim's procedure would become extremely complicated, had there been any cluster specific random effect (severity factor) with the covariates in their model. As explained by Zeger and Karim (1991) in the context of random effects generalized linear models, the likelihood involves high dimensional integration over the distribution of the random effects and as such no analytically tractable form of it can be obtained. For this basic reason, in the present article we do not follow the likelihood based approach.

Here we compare our results with that obtained by Kim. Note that apart from considering the random effects, we have taken a larger number of covariates in our analysis and hence the numerical figures of the estimates of Kim are not directly comparable with us. For comparison we consider the computations presented in Table 1 only (Γ is known). We have kept a possibility of different cut-off points (γ_{1j} 's and γ_{2j} 's) for two eyes, while Kim assumed $\gamma_{1j} = \gamma_{2j}$ for all j . Assumption of same cut-off points for the two variables may seem to be alright in the WESDR data context, but there are situations where the two variables may not be similar in nature. For example, if one considers the anxiety scores, there is reason to believe that the two sets of cut-off points are not same for overt and covert scores. Our present model allows that flexibility. The estimates of cut-off points are also not comparable with those of Kim. Kim used an arbitrary scaling of the latent variables (which is difficult to relate with the retinopathy levels), while in our analysis the scale of the latent variable is taken with due attention to the values of the retinopathy levels. This leads to the advantage of easy interpretation of our result. For example, the posterior mean of γ_{11} to be 12.44 can be easily interpreted in terms of the groupings used (10, 21-37, 43-53, 60-85). The posterior distributions of our cut-off points justify the groupings used.

Kim derived the estimate of the polychoric correlation coefficient ρ as $\hat{\rho} = 0.922$, while using our posterior means it comes out to be 0.8262. Note that Kim obtained the effect of the covariates combining the two eyes, and we have obtained them separately. The basic trend of our results remains same as those of Kim. In Kim's analysis DuD, DBP and GH have positive coefficients which is similar in our analysis for both the eyes. In our case the other regression coefficients have similar interpretation. For example, DI has positive posterior means and positive 2.5% quantile for both the eyes implying DI tends to increase retinopathy. The error variances are separately estimated for both the eyes as most part of the distribution falls in the positive domain

of the real line. The computations with some other priors for β have also been carried out. As the basic trend of the results remain same, we skip those here for the sake of brevity.

5. Analysis in the Truncated Set Up

Quite often, in some studies on bivariate ordinal data, observations corresponding to a few cells are not recorded at all. The likelihood based analysis of such truncated data becomes very much complicated as no global maximum of the likelihood function exists. This fact has been noted by Weiss (1993) in the analysis of motorcycle accident data (see Hurt et al., 1980). Hurt's data actually provides observations on two types of injuries, e.g., head or neck injury and body injury. Note that a motorcycle accident was recorded only if at least some (head/neck and/or body) injury occurred. Typically accidents with no head or neck injury and no body injury are not observed at all. The injury levels were classified on an ordinal scale between 1 and 6, reflecting its severity, using the abbreviated injury scale (AIS) (Committee on Injury Scaling, 1980), with '0' corresponds to no injury. Several covariates like helmet use (indicator), speed, alcohol use etc. were responsible for the responses. Weiss (1993) analyzed this dataset from a frequentist's viewpoint. But he rightly mentioned the problem that arises in estimating the coefficient parameters as the loglikelihood function ceases to be globally concave. In fact, if the information matrix is positive definite and the sample size is sufficiently large, the log-likelihood function will be concave in some neighborhood of the true vector of parameters and will have a local maximum in this neighborhood. Weiss (1993) described some iterative procedure and this should converge to the local maximum if the iteration is started in this neighborhood. But the difficulty remains

as an initial consistent estimate is not readily available.

Bivariate or multivariate categorical responses with a few unobserved cells arise quite often in real life problems. For example, in psychological studies, bivariate/multivariate categorical responses are usually observed for those individuals for whom the anxiety levels are significant. Also in the geographical studies relating to depth of a river bed, only the pair of measurements are recorded for which at least one measurement is significant.

Without loss of generality, let us assume that the cell $(0, 0)$ is truncated in the data. Clearly, truncation occurs as observations with both $y_{1i}^* \leq \gamma_{11}$, $y_{2i}^* \leq \gamma_{21}$ are not observed. Thus here the pair of responses (y_{1i}, y_{2i}) can take the values (j, l) where $(j, l) \in S_2 = \{(j, l) : j = 0, 1, \dots, k_1, l = 0, 1, \dots, k_2, (j, l) \neq (0, 0)\}$. Incidentally, it is quite pertinent to mention that Weiss (1993), in his frequentist probit model, did not consider the random components although these components are required to be introduced into the model as the severity factors. Then the joint distribution of y and y^* become

$$\pi(y^*, y | \beta, \Sigma, \gamma) = \prod_{i=1}^n \left[\sum_{j,l \in S_2} 1_{jl}^i 1(\gamma_{1j} < y_{1i}^* \leq \gamma_{1j+1}, \gamma_{2l} < y_{2i}^* \leq \gamma_{2l+1} | (j, l) \neq (0, 0)) \right] \times N_2(X_i \beta + Z_i b_i, \Sigma) \cdot [1 - 1(y_{1i}^* \leq \gamma_{11}, y_{2i}^* \leq \gamma_{21})]. \quad (5.1)$$

Note that here the last term within square bracket in the right hand side of (5.1) represents the truncated nature of the bivariate distribution of y_i^* . The posterior distribution of all the parameters will remain same as in section 3 except that there will be no y_i^* in (3.8) defined in the domain $\{y_{1i}^* \leq \gamma_{11}, y_{2i}^* \leq \gamma_{21}\}$. This is intuitively clear from the basic nature of the data. The Gibbs sampler can be effectively employed in this case in a similar manner.

6. Concluding Remarks

The present article is an attempt to analyze bivariate ordinal data using a random component bivariate threshold model in the presence of random severity factors. The approach is well applicable in the multivariate set up with possibly some additional difficulty in the modeling. As in the case of most of the data analysis works in the literature, here the present analysis is done assuming model validation. The examples show that if the data fits into a linear model described in section 3, the method works well. Analysis of the truncated data on motorcycle accidents could be an interesting task for better illustration of the method in the truncated set up. But, we could not access that data set in the form of accident specific raw data.

Acknowledgments: We are grateful to Drs Ronald Klein and Barbara Klein for providing the WESDR data set. The work of Drs R. Klein and B. Klein was supported by NIH grant EY 03083.

References

- Ashford, J. R. (1959). An approach to the analysis of data for semiquantitative responses in biological assay. *Biometrics* **15**, 573-581.
- Bush, C. A. and MacEachern, S. N. (1996). A semi-parametric Bayesian model for randomized block designs. *Biometrika* **83**, 275-285.
- Committee on Injury Scaling (1980). *The Abbreviated Injury Scale 1980 Revision*. Morton Grove: American Association for Automotive Medicine.

- Cox, D. R. (1970). *The analysis of binary data*. Chapman and Hall, London.
- Dale, J. R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics* **42**, 909-917.
- Doss, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *Ann. Statist.* **22**, 1763-1786.
- Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Statist. Assoc.* **89**, 268-277.
- Ferguson, T. S. (1973). A Bayesian analysis of some non-parametric problems. *Ann. Statist.* **1**, 209-230.
- 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. *J. Amer. Statist. Assoc.* **85**, 972-985.
- Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **6**, 721-741.
- Hurt Jr., H. H., Ouellet, J. V. and Thom, D. R. (1981). Motorcycle accident cause factors and identification of counter-measures, Vol I. Technical Report. Washington DC: US Department of Transportation.
- Kim, K. (1995). A bivariate cumulative probit regression model for ordered categorical data. *Statistics in Medicine* **14**, 1341-1352.
- Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D. and DeMets, D. L. (1984). The Wisconsin Epidemiologic study of diabetic retinopathy, II: prevalence and

- risk of diabetic retinopathy when age at diagnosis is less than 30 years. *Arch. Ophthalmol.* **102**, 520-526.
- Krugg, S. E., Scheier, I. H. and Cattell, R. B. (1976). *Handbook for the IPAT anxiety scale*. Institute for personality and ability testing, Illinois.
- Lesaffre, E. and Molenberghs, G. (1991). Multivariate probit analysis: a neglected procedure in medical statistics. *Statistics in Medicine* **10**, 1391-1403.
- Liu, J. (1996). Nonparametric hierarchical Bayes via sequential imputation. *Ann. Statist.* **24**, 911-930.
- MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics* **23**, 727-741.
- MacEachern, S. N. and Muller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*.
- McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Ser. B* **42**, 109-142.
- Molenberghs, G. and Lesaffre, E. (1994). Marginal modeling of correlated ordinal data using a multivariate Plackett distribution. *Journal of the American Statistical Association* **89**, 633-644.
- Mukhopadhyay, S. (1989). Working status and stress of middle class women in Calcutta. *Journal of Biosocial Science* **21**, 109-114.
- Muller, P., Erkanli, A. and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**, 67-79.

- Muthen, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* **49**, 115-132.
- Snell, E. J. (1964). A scaling procedure for ordered categorical data. *Biometrics* **20**, 592-607.
- Weiss, A. A. (1993). A bivariate ordered probit model with truncation: helmet use and motorcycle injuries. *Applied Statistics* **42**, 487-499.
- West, M., Muller, P. and Escobar, M. D. (1994). Hierarchical priors and mixture models, with applications in regression and density estimation. In *Aspects of Uncertainty: A Tribute to D. V. Lindley*, A. F. M. Smith and P. R. Freeman (eds). London: Wiley.
- Wilks, W. R., Wang, C. C., Yvonnet, B. and Coursaget, P. (1993). Random effects models for longitudinal data using Gibbs sampling. *Biometrics* **49**, 441-453.
- Williamson, J. M., Kim, K. and Lipsitz, S. R. (1995). Analyzing bivariate ordinal data using a global odds ratio. *Journal of the American Statistical Association* **90**, 1432-1437.
- Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association* **86**, 79-86.

Table 1: Posterior summary statistics for noninformative prior.

eye	node	mean	s.e.	MC error	2.5%	median	97.5%
left	Constant	0.0212	0.469	0.1134	-0.561	0.534	1.729
	RS	4.128	0.1379	0.02431	3.82	4.15	4.556
	left ME	2.333	1.219	0.164	0.1576	2.398	4.887
	left RE	-0.0555	0.103	0.0088	-0.3075	-0.08065	0.1234
	left IOP	0.1226	0.0444	0.007814	0.0357	0.1366	0.1976
	AgD	0.0127	0.02179	0.00558	-0.00123	0.02718	0.0544
	DuD	0.07023	0.0341	0.004422	0.02356	0.07375	0.1109
	GH	0.04922	0.03677	0.01669	0.02754	0.05637	0.1116
	SBP	-0.03346	0.01723	0.002224	-0.05429	-0.02567	-0.0602
	DBP	0.05715	0.03169	0.005054	0.009234	0.02674	0.07978
	BMI	0.12282	0.03326	0.00508	0.00806	0.12394	0.16199
	PR	0.0940	0.01833	0.003004	0.02524	0.0778	0.1297
	Sex	0.4012	0.0314	0.0121	0.3180	0.3914	0.4728
	UP	0.6094	0.1607	0.1366	0.1736	0.6229	0.8688
	DI	0.4488	0.1433	0.0902	0.0842	0.4237	0.6949
	AR	1.156	0.7012	0.1198	-0.05947	0.9972	2.36
right	Constant	0.0406	0.341	0.108	-0.663	0.674	1.876
	RS	4.235	0.112	0.01934	3.654	4.209	4.578
	right ME	2.455	1.127	0.1566	0.1635	2.456	4.997
	right RE	-0.1084	0.0976	0.0089	-0.2876	-0.0913	0.1314
	right IOP	0.1308	0.0432	0.007677	0.0555	0.1448	0.2105
	AgD	0.0144	0.02287	0.00439	-0.00521	0.02574	0.0601
	DuD	0.06653	0.0321	0.006264	0.02498	0.0811	0.1137
	GH	0.0534	0.02254	0.01102	0.02867	0.07654	0.1239
	SBP	-0.03826	0.01055	0.002044	-0.0542	-0.02913	-0.004449
	DBP	0.0149	0.02742	0.004886	0.00382	0.020446	0.04936
	BMI	0.11302	0.04033	0.006129	0.008555	0.0928	0.1748
	PR	0.0874	0.02302	0.004005	0.02319	0.07899	0.12625
	Sex	0.3874	0.0841	0.00933	0.3262	0.3801	0.4726
	UP	0.7285	0.1342	0.1055	0.2425	0.7099	0.9214
	DI	0.5052	0.1366	0.0744	0.1244	0.4977	0.7432
	AR	1.123	0.3672	0.06494	-0.0464	0.9742	2.2485
left	γ_{11}	12.44	0.349	0.05008	12.21	13.62	14.4
	γ_{12}	40.58	0.643	0.1056	38.55	40.24	41.52
	γ_{13}	62.68	1.167	0.2064	57.69	61.45	64.28
right	γ_{21}	13.66	0.4786	0.0828	12.88	14.13	15.92
	γ_{22}	39.48	0.5834	0.1019	38.24	39.88	41.65
	γ_{23}	60.44	0.996	0.1247	58.02	61.05	64.66
	σ_{11}	5.48	0.723	0.1113	3.57	5.12	6.08
	σ_{12}	3.54	0.342	0.0513	2.53	3.21	4.16
	σ_{22}	3.35	0.318	0.0451	2.27	3.21	4.165