A Nonparametric Bayesian Model for Multivariate Ordinal Data

Athanasios Kottas, University of California at Santa Cruz
Peter Müller, The University of Texas M. D. Anderson Cancer Center
Fernando A. Quintana, Pontificia Universidad Católica de Chile
Fernando A. Quintana, Departamento de Estadística, Casilla 306, Santiago 22, Chile

Key Words: Contingency tables; Dirichlet process; Markov chain Monte Carlo; polychoric correlations.

Abstract:

We propose a new approach to inference for ordinal outcomes of dimension k. This is done by means of a mixture of multivariate probits model on latent scores. The model can be used for the cell probabilities of any k-dimensional contingency table. Our method achieves two main goals. First, the model allows arbitrarily accurate approximation of any given set of probabilities on the outcomes. And secondly, the proposed model allows, without loss of generality, to fix the cutoffs of the latent scores, avoiding the need for complicated posterior simulation of these cutoffs.

1. Introduction

Consider measurements on the values of k ordinal categorical variables $V_1, \ldots, V_k$ for each of n subjects. Let $C_j \geq 2$ denote the number of categories for the jth variable, $j = 1, \ldots, k$, and let $n_{\ell_1 \cdots \ell_k}$ be the number of observations with $\boldsymbol{V} = (V_1, \ldots, V_k) = (\ell_1, \ldots, \ell_k)$. Denote by $p_{\ell_1 \cdots \ell_k} = P(V_1 = \ell_1, \ldots, V_k = \ell_k)$ the classification probability for the $(\ell_1, \ldots, \ell_k)$ cell. These data can be summarized in a multidimensional contingency table with $C = \prod_{j=1}^{k} C_j$ cells, and frequencies $\{n_{\ell_1 \cdots \ell_k}\}$ subject to $\sum_{\ell_1 \cdots \ell_k} n_{\ell_1 \cdots \ell_k} = n$.

The data structure just described arises in many applications, which explains the vast statistical literature available. Many examples, applications and technical details can be found in Bishop, Fienberg, and Holland (1975), Goodman (1985), Read and Cressie (1988) and references therein. Log-linear models are a popular choice for the analysis of such examples. However, the number of parameters grows quickly with C, which gives rise to several problems related to interpretation, prior elicitation and assessment of association between categorical variables.

An alternative modeling strategy consists of introducing latent variables. Examples include Albert and Chib (1993), Cowles, Carlin, and Connett (1996), Chib and Greenberg (1998), Bradlow and Zaslavsky (1999), Chen and Dey (2000) and Chib (2000) for ordinal regression models, Johnson and Albert (1999) for the analysis of data from multiple raters and Newton, Czado, and Chappell (1995) for semiparametric binary regression. This approach requires the definition of cutoffs $-\infty = \gamma_{j,0} < \gamma_{j,1} < \cdots < \gamma_{j,C_j-1} < \gamma_{j,C_j} = \infty$, for each $j = 1, \ldots, k$, and a k-dimensional latent variable vector $\boldsymbol{Z} = (Z_1, \ldots, Z_k)$ such that for all $\ell_1, \ldots, \ell_k$

$$p_{\ell_1 \cdots \ell_k} = P\left(\bigcap_{j=1}^{k} \left\{\gamma_{j,\ell_j-1} < Z_j \leq \gamma_{j,\ell_j}\right\}\right). \quad (1)$$

A common assumption is $\boldsymbol{Z} \sim N_k(\boldsymbol{0}, \boldsymbol{S})$, i.e., a k-dimensional normal distribution. One advantage of this model is the parsimony compared to the saturated log-linear model. In addition, $\rho_{st} = \text{Corr}(Z_s, Z_t) = 0$, $s \neq t$, implies independence of the corresponding categorical variables. The coefficients ρ_{st}, $s \neq t$, are known as *polychoric correlation coefficients* and have been extensively used in the social sciences as a measure of the association between pairs of the (observed) categorical variables. See for instance, Olsson (1979) and more recently, Ronning and Kukuk (1996) and references therein.

However, the above model is not an appropriate choice for contingency tables that concentrate most of the data near the borders or corners and are rather sparse in the central cells. Also, the multivariate normal probit model implicitly assumes that the same polychoric correlations are globally valid. It does not allow for the polychoric correlations to vary across the contingency table.

As an illustration of the above issue, consider $n = 100$ latent scores simulated from a mixture of two bivariate normal distributions, with means $(-1.5, -1.5)^T$ and $(0.5, 0.5)^T$, marginal variances all equal to 0.25, and correlations -0.7 and 0.35, respectively. Setting cutoffs -2.5, -1.5, -0.5, 0.5, and 1.5 for both dimensions, the simulated scores imply a 6×6 contingency table. Consider now the model where the likelihood is induced by (1), $\boldsymbol{Z} \sim N_2(\boldsymbol{m}, \boldsymbol{S})$, $\boldsymbol{m} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, completed with conjugate-style priors for $\boldsymbol{\mu}$, $\boldsymbol{S}$, and $\boldsymbol{\Sigma}$, and vague hyperparameter choices. As an obvious competing model we consider the two-component mixture $\boldsymbol{Z} \sim pN_2(\boldsymbol{m}_1, \boldsymbol{S}_1) + (1-p)N_2(\boldsymbol{m}_2, \boldsymbol{S}_2)$, with analogous prior structure on all parameters. Figure 1 shows

Quintana's research was partially supported by grant FONDECYT 1020712

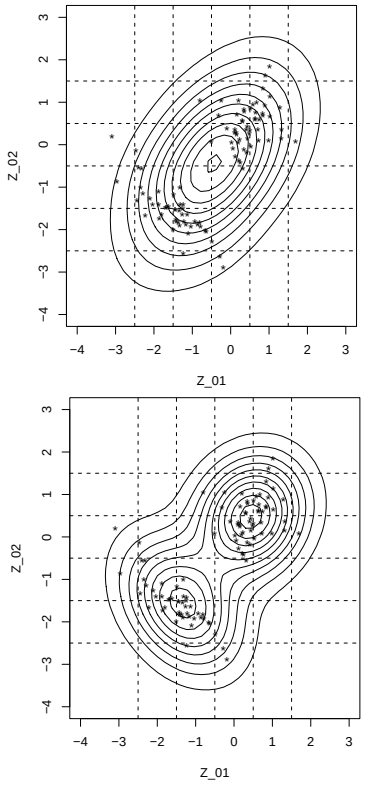

Figure 1: Posterior predictive distribution $p(\boldsymbol{Z}_0|\boldsymbol{V})$ together with simulated latent scores. Top panel shows model with one mixture component, while bottom panel represents model with two mixture components.

contour plots of the posterior predictive distribution for a new $\boldsymbol{Z}_0$ under each model. Clearly, the one-component model does not perform well in this specific example. In particular, the central area, where no data was recorded, is forced to receive the highest probability mass under this model. As expected, the two-component model provides a much better fit to the data, as by construction, it captures the two groups of latent scores.

The example just shown immediately raises the question: how many mixture components should we use? In practice, we do not know that number. On the other hand, the cutoffs were arbitrarily fixed to coincide with the ones used to simulate latent scores. A probability model is required for the cutoffs $\{\gamma_{j,\ell_j}\}$ if they are to be treated as unknown quantities. And otherwise, how should we choose them? Such considerations motivate the use of more flexible families of distributions for $\boldsymbol{Z}$. The literature on fully Bayesian inference in this context is rather scarce. This appears to be limited to Albert (1992), involving bivariate log-normal and t distributions, and Chen and Dey (2000), who discuss scale mixtures of multivariate normals. An additional computational complication arises when resampling the $\{\gamma_{j,\ell_j}\}$s. Indeed, the values of $\boldsymbol{Z}$ can become tightly clustered around a given γ_{j,ℓ_j} yielding a nearly degenerate full conditional. Johnson and Albert (1999) handle this problem via hybrid MCMC samplers.

We present next a model that is flexible enough to represent essentially any set $\{p_{\ell_1,\ldots,\ell_k}\}$ for cell probabilities. At the same time, we argue that without loss of generality the cutoffs for the latent variables can be arbitrarily fixed, for example on an equally spaced grid. Thus our model provides the most general approach for the analysis of contingency tables, and at the same time is easier to implement than other existing Bayesian methods.

2. A Nonparametric Model

Consider n vectors of ordinal categorical variables $\boldsymbol{V}_i = (V_{i1}, \ldots, V_{ik})$, $i = 1, \ldots, n$. Corresponding to each $\boldsymbol{V}_i$ there is a vector of latent variables $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ik})$, $i = 1, \ldots, n$, linked with the ordinal outcomes as $V_i = \ell_j$ if and only if $\gamma_{j,\ell_{j-1}} < Z_j \leq \gamma_{j,\ell_{j-1}}$, thus defining cell probabilities as in (1). We denote this deterministic link by $p(\boldsymbol{V}|\boldsymbol{Z})$.

Our approach considers now a mixture of normals model for the $\boldsymbol{Z}_i$s, with respect to both location and covariance matrix. Letting $\boldsymbol{\theta} = (\boldsymbol{m}, \boldsymbol{S})$ we assume

$$\boldsymbol{Z}_i \overset{iid}{\sim} f, \text{ with } f(\cdot|G) = \int p_{N_k}(\cdot|\boldsymbol{m}, \boldsymbol{S})\, dG(\boldsymbol{m}, \boldsymbol{S}), \tag{2}$$

i.e., a mixture of normals, with respect to a mixing measure G. We model the random mixing measure G as

$$G|\alpha, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \boldsymbol{D} \sim \mathcal{D}(\alpha G_0), \tag{3}$$

a Dirichlet Process (DP) prior (Ferguson, 1973) with total mass parameter α and baseline distribution G_0. Furthermore, we assume G_0 is specified as follows: $G_0(\boldsymbol{m}, \boldsymbol{S}) \equiv N_k(\boldsymbol{m}|\boldsymbol{\lambda}, \boldsymbol{\Sigma})$ IWish$_k(\boldsymbol{S}|\nu, \boldsymbol{D})$, where $N_k(\boldsymbol{x}|\ldots)$ and IWish$_k(\boldsymbol{A}|\ldots)$ denote, respectively, a k-dimensional normal distribution for $\boldsymbol{x}$ and an inverse Wishart distribution for the $k \times k$ random matrix $\boldsymbol{A}$. Introducing latent variables $\boldsymbol{\theta}_i = (\boldsymbol{m}_i, \boldsymbol{S}_i)$, we can equivalently rewrite (2) as a hierarchical model as

$$\boldsymbol{Z}_i|\boldsymbol{\theta}_i \overset{ind}{\sim} N_k(\boldsymbol{m}_i, \boldsymbol{S}_i), \qquad i = 1, \ldots, n, \tag{4}$$

where, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ are a sample from the mixing distribution G,

$$\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n|G \overset{iid}{\sim} G. \tag{5}$$

The representation (2) highlights the interpretation of the model as a mixture for the latent variables. In addition, the model is stated in terms of covariance rather than correlation matrices, which avoids a number of associated difficulties. See, for example, Chib and Greenberg (1998), Daniels and Kass (1999) or McCulloch, Polson, and Rossi (2000). Also, the variances on the diag-

onal of $\boldsymbol{S}_i$ play an important role in defining the mixture. Smaller variances imply that the corresponding polychoric correlations are only locally valid, for ordinal scores close to $\boldsymbol{m}_i$.

To complete the model specification, we assume independent hyperpriors as follows:

$$\begin{array}{rclrcl} \alpha & \sim & \text{Gamma}(a_0, b_0) & \boldsymbol{\lambda} & \sim & N_k(\boldsymbol{q}, \boldsymbol{Q}) \\ \boldsymbol{\Sigma} & \sim & \text{IWish}_k(b, \boldsymbol{B}) & \boldsymbol{D} & \sim & \text{Wish}_k(c, \boldsymbol{C}), \end{array}$$

with fixed scalar hyperparameters ν, a_0, b_0, b, c, a k-dimensional vector $\boldsymbol{q}$, and $k \times k$ positive definite matrices $\boldsymbol{Q}$, $\boldsymbol{B}$ and $\boldsymbol{C}$.

MCMC schemes involving the DP are straightforwardly implemented, which explains its enormous popularity. In fact, computational complexity is, theoretically, independent of the dimension of $\boldsymbol{\theta}$. A key property of the DP for this application, is its discrete behavior (e.g., Blackwell and MacQueen, 1973; Sethuraman, 1994). Denoting by δ_x a point mass at x, the DP can be represented as

$$G = \sum_{h=1}^{\infty} w_h \delta_{\boldsymbol{\theta}_h} \tag{6}$$

with stochastically ordered weights w_h. See Sethuraman (1994) for details. Each different point mass $\boldsymbol{\theta}_h$ in the mixture (6) thus represents a different polychoric correlation (induced by $\boldsymbol{S}_i$), located at the factor levels of the table as specified by $\boldsymbol{m}_i$. In addition, the DP can be calibrated in such a way that *a priori* only few different polychoric correlations are considered. *A posteriori*, the number of weights and polychoric correlations spread through the table is determined as supported by the data. In terms of (5), the discreteness just described implies that there may be ties among the latent parameters $\{\boldsymbol{\theta}_i\}_{i=1}^n$. This immediately induces a *clustering structure* by grouping together all those components that have identical $\boldsymbol{\theta}$ value.

The nonparametric mixture model $f(\cdot|G)$ has another key property. From (6) it follows that $f(\cdot|G)$ can be written as a countable mixture of normal kernels, $\sum_{h=1}^{\infty} w_h p_{N_k}(\cdot|\boldsymbol{m}_h, \boldsymbol{S}_h)$. Then, given any set of table cell probabilities $\{p_{\ell_1 \ldots \ell_k}\}$ for the contingency table we can think of a mixture with weights $\{w_h\}$ matched to the desired cell probabilities $\{p_{\ell_1,\ldots,\ell_k}\}$, i.e., $w_h = p_{\ell_1,\ldots,\ell_k}$ for some h. Next, we choose $(\boldsymbol{m}_h, \boldsymbol{S}_h)$ such that $(1-\epsilon)$ of the distribution $N(\boldsymbol{m}_h, \boldsymbol{S}_h)$ falls within the rectangle $\prod_{i=1}^k (\gamma_{j,\ell_j - 1}, \gamma_{j,\ell_j}]$. This induces via the deterministic link model $p(\boldsymbol{V}|\boldsymbol{Z})$ a distribution on the table cell probabilities that is arbitrarily close to $\{p_{\ell_1 \ldots \ell_k}\}$. Therefore, the mixture model can accommodate any given set of contingency table probabilities, including "irregular patterns" that can not be explained by a multivariate normal probit model with a global polychoric correlation that applies across the entire table. In addition, the above can be argued to be entirely independent of the chosen set of cutoffs. Hence, there is no loss of generality in assuming them to be fixed when implementing the model.

3. Illustration with a Data Set of Interrater Agreement

We consider a data set from Melia and Diener-West (1994) reporting extent of scleral extension (extent to which a tumor has invaded the sclera or "white of the eye") as coded by two raters, A and B, for each of $n = 885$ eyes. The coding scheme uses five categories, ranging from least to most severe: 1 for "none or innermost layers", 2 for "within sclera, but does not extend to scleral surface", 3 for "extends to scleral surface", 4 for "extrascleral extension without transection" and 5 for "extrascleral extension with presumed residual tumor in the orbit". The data set is available from the StatLib data sets archive at http://lib.stat.cmu.edu/datasets/csb/ch16a.dat. We provide the observed cell frequencies in Table 1.

To fit the model to these data, we used cutoffs -1,0,1,2 for both variables. For the hyperparameters we chose the following values: $\boldsymbol{q} = (0,0)^T$, $\boldsymbol{H} = \text{diag}(6.25, 6.25)$, $\boldsymbol{Q} = \boldsymbol{B} = \boldsymbol{H}$ and $b = 4$ for the priors for $\boldsymbol{\lambda}$ and $\boldsymbol{\Sigma}$. For the total mass parameter we chose a $\text{Gamma}(2, 0.9)$ distribution. Finally, we set $\nu = 10$, $c = 5$ and $\boldsymbol{C} = \text{diag}(8.75, 8.75)$ yielding a quite dispersed prior for $\boldsymbol{D}$. We found that the posterior distributions of $\boldsymbol{\lambda}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{D}$ were quite concentrated compared with their priors, indicating that these prior choices are vague compared to the likelihood.

Updating the latent mixture parameters $\{\boldsymbol{\theta}_i\}_{i=1}^n$ and the hyperparameters $\boldsymbol{\lambda}$, $\boldsymbol{\Sigma}$, $\boldsymbol{D}$ and α is accomplished via standard posterior simulation methods for DP mixtures. See, for example, MacEachern and Müller (1998).

The mixture model identifies four components that are important in describing clustering in terms of factor levels in the table and polychoric correlations. We illustrate this clustering in Figure 2. Figure 2(a) contains 2500 draws from the posterior predictive distribution $p(\boldsymbol{Z}_0|\text{data})$ and Figure 2(b) shows one draw from the posterior when the number of different point masses is 4. Here, the 4 ellipsoids are determined from the corresponding (common) means and covariance matrices of the clusters. The associated posterior weights, proportional to the size of the clusters, are also indicated in the plot. The posterior of the number of imputed clusters is given in Figure 3(d). Note that all the posterior mass is between 4 and 15, with median 7. However, all posterior draws with $n^* > 4$ have very small weights for clusters other than the four depicted in Figure 2.

Figure 3(a) shows the posterior means and the .025 and .975 posterior percentiles for ρ_i, $i = 1, \ldots, 885$. The analogous plots for $\boldsymbol{S}_{i,11}$ and $\boldsymbol{S}_{i,22}$, $i = 1, \ldots, 885$, are

Table 1: Observed cell frequencies (in bold) and posterior summaries for table cell probabilities. Rows correspond to rater A and columns to rater B.

	1	2	3	4	5
1	**.3288** .3261	**.0836** .0869	**.0011** .0013	**.0011** .0020	**.0011** .0007
	(.2945, .3590)	(.0696, .1078)	(.0001, .0044)	(.0003, .0056)	(.0000, .0027)
2	**.2102** .2135	**.2893** .2826	**.0079** .0082	**.0079** .0069	**.0034** .0031
	(.1858, .2428)	(.2521, .3141)	(.0031, .0152)	(.0022, .0143)	(.0007, .0075)
3	**.0023** .0023	**.0045** .0055	**.0000** .0016	**.0023** .0022	**.0000** .0008
	(.0004, .0065)	(.0017, .0107)	(.0003, .0038)	(.0004, .0062)	(.0000, .0032)
4	**.0034** .0042	**.0113** .0102	**.0011** .0023	**.0158** .0143	**.0023** .0028
	(.0012, .0094)	(.0042, .0187)	(.0004, .0060)	(.0066, .0240)	(.0006, .0069)
5	**.0011** .0012	**.0079** .0071	**.0011** .0019	**.0090** .0083	**.0034** .0039
	(.0001, .0041)	(.0026, .0140)	(.0003, .0054)	(.0034, .0153)	(.0009, .0090)

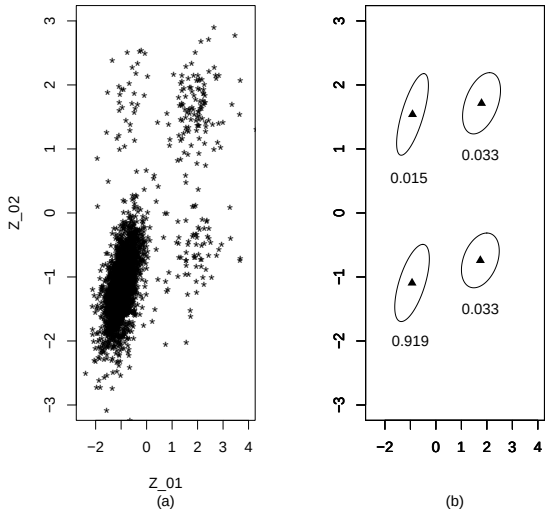

Figure 2: Panel (a) plots draws from $p(\boldsymbol{Z}_0|\text{data})$. Panel (b) shows one posterior draw for means and covariances when the number of clusters is 4. The corresponding weights – proportional to cluster sizes – are indicated below each ellipsoid.

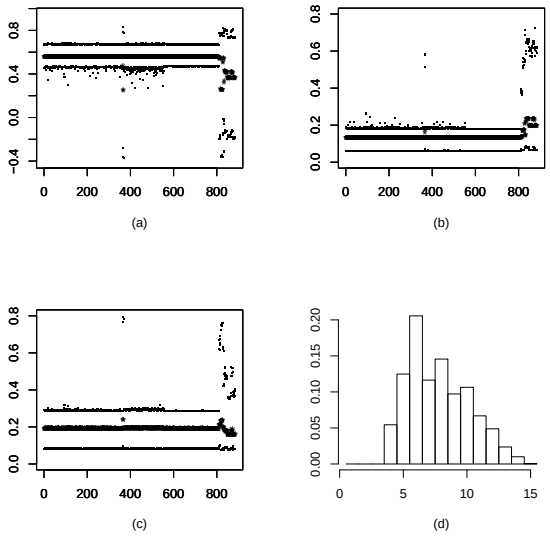

Figure 3: Panels (a), (b) and (c) plot, for $i = 1, \ldots, 885$, posterior summaries for ρ_i, $\text{Var}(Z_{i1})$ and $\text{Var}(Z_{i2})$, respectively. Panel (d) provides the posterior for number of clusters.

provided in Figures 3(b) and 3(c), respectively. These plots also capture the variability in the corresponding posteriors.

These results clearly show the utility of the proposed mixture model for this particular data set. Although one of the clusters clearly dominates the others, identifying the other three is quite important. One of them quantifies agreement for large values (4 and 5) in the coding scheme, whereas the other two indicate regions of the table where the two raters tend to disagree.

4. Discussion

The basis of our approach to multivariate ordinal data analysis is a DP mixture model for latent variables defining classification probabilities in the resulting contingency table. The model has two key features. First, the flexibility provided by the probability model on latent variables allows us to handle virtually any data structure. Second, this flexibility can be achieved with fixed cutoffs, thus avoiding the most difficult computational challenge arising in posterior simulation for related models. The example illustrates these points.

References

Albert, J. H. (1992), "Bayesian Estimation of the Polychoric Correlation Coefficient," *Journal of Statistical Computation and Simulation*, 44, 47–61.

Albert, J. H. and Chib, S. (1993), "Bayesian Analysis of

Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: M.I.T. Press.

Blackwell, D. and MacQueen, J. B. (1973), "Ferguson distributions via Pólya urn schemes," *The Annals of Statistics*, 1, 353–355.

Bradlow, E. T. and Zaslavsky, A. M. (1999), "Hierarchical Latent Variable Model for Ordinal Data From a Customer Satisfaction Survey With "No Answer" Responses," *Journal of the American Statistical Association*, 94, 43–52.

Chen, M.-H. and Dey, D. K. (2000), "Bayesian Analysis for Correlated Ordinal Data Models," in *Generalized Linear Models: A Bayesian Perspective*, eds. D. K. Dey, W. S. Ghosh, and B. Mallick, 135–162, New York: Marcel Dekker.

Chib, S. (2000), "Bayesian Methods for Correlated Binary Data," in *Generalized Linear Models: A Bayesian Perspective*, eds. D. K. Dey, W. S. Ghosh, and B. Mallick, 113–131, New York: Marcel Dekker.

Chib, S. and Greenberg, E. (1998), "Analysis of multivariate probit models," *Biometrika*, 85, 347–361.

Cowles, M. K., Carlin, B. P., and Connett, J. E. (1996), "Bayesian Tobit Modeling of Longitudinal Ordinal Clinical Trial Compliance Data With Nonignorable Missingness," *Journal of the American Statistical Association*, 91, 86–98.

Daniels, M. J. and Kass, R. E. (1999), "Nonconjugate Bayesian Estimation of Covariance Matrices and Its Use in Hierarchical Models," *Journal of the American Statistical Association*, 94, 1254–1263.

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

Goodman, L. A. (1985), "The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries," *The Annals of Statistics*, 3, 10–69, Rietz Memorial Lecture.

Johnson, V. E. and Albert, J. H. (1999), *Ordinal Data Modeling*, New York: Springer.

MacEachern, S. N. and Müller, P. (1998), "Estimating Mixture of Dirichlet Process Models," *Journal of Computational and Graphical Statistics*, 7, 223–338.

McCulloch, R. E., Polson, N. G., and Rossi, P. E. (2000), "A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters," *Journal of Econometrics*, 99, 173–193.

Melia, B. M. and Diener-West, M. (1994), "Modeling interrater agreement on an ordered categorical scale," in *Case Studies in Biometry*, eds. N. Lange, L. Ryan, L. Billard, D. Brillinger, L. Conquest, and J. Greenhouse, 323–338, John Wiley and Sons, New York.

Newton, M. A., Czado, C., and Chappell, R. (1995), "Bayesian Inference for Semiparametric Binary Regression," *Journal of the American Statistical Association*, 91, 132–141.

Olsson, U. (1979), "Maximum likelihood estimation of the polychoric correlation coefficient," *Psychometrika*, 44, 443–460.

Read, T. R. C. and Cressie, N. A. C. (1988), *Goodness-of-fit Statistics for Discrete Multivariate Data*, New York: Springer.

Ronning, G. and Kukuk, M. (1996), "Efficient estimation of ordered probit models," *Journal of the American Statistical Association*, 91, 1120–1129.

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