

Nonparametric Bayesian Modeling for Multivariate Ordinal Data

Athanasios Kottas, Peter Müller and Fernando Quintana*

Abstract

We propose a probability model for k -dimensional ordinal outcomes, i.e., we consider inference for data recorded in k -dimensional contingency tables with ordinal factors. The proposed approach is based on full posterior inference, assuming a flexible underlying prior probability model for the contingency table cell probabilities. We use a variation of the traditional multivariate probit model, with latent scores that determine the observed data. In our model, a mixture of normals prior replaces the usual single multivariate normal model for the latent variables. By augmenting the prior model to a mixture of normals we generalize inference in two important ways. First, we allow for different polychoric correlation coefficients across the contingency table. Second, inference in ordinal multivariate probit models is plagued by problems related to the choice and resampling of cutoffs defined for these latent variables. We show how the proposed mixture model approach entirely removes these problems. We illustrate the methodology with two examples, one simulated data set and one data set of interrater agreement.

*Athanasios Kottas is Assistant Professor, Department of Applied Mathematics and Statistics, Baskin School of Engineering, University of California at Santa Cruz, Santa Cruz, CA 95064 (E-mail: thanos@ams.ucsc.edu). Peter Müller is Professor, Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030 (E-mail: pm@odin.mdacc.tmc.edu). Fernando Quintana is Associate Professor, Departamento de Estadística, Pontificia Universidad Católica de Chile, Avenida Vicuña Mackenna 4860, Santiago, CHILE (E-mail: quintana@mat.puc.cl). Quintana's research was partially supported by grant FONDECYT 1020712. The order of author names is strictly alphabetical.

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

1 Introduction

We consider inference for k -dimensional ordinal outcomes. We define a mixture of multivariate probits model that can represent any set of k -dimensional contingency table cell probabilities. The proposed approach generalizes the traditional multivariate probit model, and at the same time allows for significant simplification of computational complexity. Computational simplicity is achieved by avoiding the need to impute cutoffs for the latent scores. Increased modeling flexibility is provided by allowing arbitrarily accurate approximation of any given set of probabilities on the outcomes. In particular, polychoric correlations (i.e., correlations of the latent scores) are allowed to vary across the contingency table.

Assume that for each of n experimental units the values of k ordinal categorical variables V_1, \dots, V_k are recorded. Let $C_j \geq 2$ represent the number of categories for the j th variable, $j = 1, \dots, k$, and denote by $n_{\ell_1 \dots \ell_k}$ the number of observations with $\mathbf{V} = (V_1, \dots, V_k) = (\ell_1, \dots, \ell_k)$. Denote by $p_{\ell_1 \dots \ell_k} = P(V_1 = \ell_1, \dots, V_k = \ell_k)$ the classification probability for the (ℓ_1, \dots, ℓ_k) cell. The data can be summarized in a multidimensional contingency table with $C = \prod_{j=1}^k C_j$ cells, with frequencies $\{n_{\ell_1 \dots \ell_k}\}$ constrained by $\sum_{\ell_1 \dots \ell_k} n_{\ell_1 \dots \ell_k} = n$.

Inference for such data structures is of interest in many applications. The related statistical literature is correspondingly diverse and extensive. Many examples, applications and technical details can be found in Bishop et al. (1975), Goodman (1985), Read and Cressie (1988) and references therein. Log-linear models are a popular choice for the analysis of this data structure. However, the typically large number of parameters gives rise to a number of difficulties related to interpretation, prior elicitation and assessment of association between categorical variables.

An alternative modeling strategy involves the introduction of latent variables. Examples include Albert and Chib (1993), Cowles et al. (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 et al.

(1995) for semiparametric binary regression. The common idea in these approaches is to introduce cutoffs $-\infty = \gamma_{j,0} < \gamma_{j,1} < \dots < \gamma_{j,C_j-1} < \gamma_{j,C_j} = \infty$, for each $j = 1, \dots, k$, and a k -dimensional latent variable vector $\mathbf{Z} = (Z_1, \dots, Z_k)$ such that for all ℓ_1, \dots, ℓ_k

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

A common distributional assumption is $\mathbf{Z} \sim N_k(\mathbf{0}, \mathbf{S})$, 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 are traditionally 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 described 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. A typical example where such heterogeneity might arise is interrater agreement data. Different raters might agree about extremely high or low scores. But there might be considerable disagreement about scoring average observations. These limitations motivate the introduction of more flexible families of distributions for the latent variables \mathbf{Z} . Literature on fully Bayesian inference in this context is rather limited. We are only aware of Albert (1992), involving bivariate log-normal and t distributions, and Chen and Dey (2000), where certain scale mixtures of multivariate normals are considered.

An important practical issue is related to the choice of the cutoffs γ_{j,ℓ_j} . First, identifiability constraints complicate inference. Second, if the cutoffs are considered unknown parameters, inference is complicated by the fact that they are highly correlated with the latent variables \mathbf{Z} . In particular, when abundant data are available, the values of \mathbf{Z} can become tightly clustered around a given γ_{j,ℓ_j} leaving little room for the cutoff to move when implementing a Markov chain Monte Carlo (MCMC) posterior simulation scheme. In this

case, the corresponding full conditional becomes nearly degenerate. Johnson and Albert (1999) handle this problem via hybrid MCMC samplers.

In this article we propose a nonparametric probability model for the latent variables \mathbf{Z} employing a Dirichlet process mixture of normals prior. We show that this provides the required flexibility to accommodate virtually any desired pattern in k -dimensional contingency tables. At the same time we argue that, under the proposed model, we can without loss of generality fix the cutoffs. Therefore, from a modeling and inferential perspective, we provide the most general framework for the analysis of contingency tables, while, from a practical perspective, we provide an approach that is easier to implement than existing Bayesian methods.

The rest of this article is organized as follows. Section 2 states our model, discussing its main features. Section 3 discusses simulation-based model fitting and posterior predictive inference. The methods are illustrated with two examples in section 4. We conclude with additional comments and discussion in section 5.

2 A Nonparametric Modeling Approach to Polychoric Correlations

2.1 The Model

We define a model for n vectors of ordinal categorical variables $\mathbf{V}_i = (V_{i1}, \dots, V_{ik})$, $i = 1, \dots, n$. First, as in (1), we introduce latent variables $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{ik})$, $i = 1, \dots, n$, such that

$$p(\mathbf{V}_i | \mathbf{Z}_i) = \sum_{\ell_1, \dots, \ell_k} \left\{ I(V_{i1} = \ell_1, \dots, V_{ik} = \ell_k) I \left\{ \bigcap_{j=1}^k \{ \gamma_{j, \ell_j - 1} < Z_{ij} \leq \gamma_{j, \ell_j} \} \right\} \right\}, \quad (2)$$

where $I(A)$ denotes the indicator function of set A and $p(\mathbf{V} | \mathbf{Z})$ denotes the link between the categorical variables and the latent continuous variables. Equation (2) describes a deterministic mapping of \mathbf{Z} to \mathbf{V} . We choose to use notation as a (degenerate) probability distribution $p(\mathbf{V} | \mathbf{Z})$ because this simplifies notation in the upcoming discussion. An important feature of the proposed model is that it allows us to use fixed cutoffs γ_{j, ℓ_j} . See the

discussion towards the end of this subsection.

Modeling proceeds now with the k -dimensional latent vectors \mathbf{Z}_i . We generalize traditional multivariate normal models by assuming a mixture of normals model. The mixture is with respect to both, location \mathbf{m} and covariance matrix \mathbf{S} of the normal kernel. We define a probability model for the mixture distribution by assuming a prior probability model for the mixing distribution $G(\mathbf{m}, \mathbf{S})$. For reasons of technical convenience and interpretability we choose a Dirichlet process (DP) prior (Ferguson, 1973). The discrete nature of the DP simplifies interpretation. Each point mass (\mathbf{m}, \mathbf{S}) in the discrete mixing distribution G corresponds to a different polychoric correlation, implicit in \mathbf{S} . The location \mathbf{m} specifies the factor levels of the contingency table where \mathbf{S} defines the polychoric correlation. Details of the construction are discussed below.

We assume $\mathbf{Z}_i \stackrel{iid}{\sim} f$, with $f(\cdot|G) = \int p_{N_k}(\cdot|\mathbf{m}, \mathbf{S}) dG(\mathbf{m}, \mathbf{S})$. Here, $p_{N_k}(\cdot|\mathbf{m}, \mathbf{S})$ denotes the density of a $N_k(\mathbf{m}, \mathbf{S})$ distribution. The mixture model f can be equivalently written as a hierarchical model by introducing latent variables $\boldsymbol{\theta}_i = (\mathbf{m}_i, \mathbf{S}_i)$ and breaking the mixture as

$$\mathbf{Z}_i|\boldsymbol{\theta}_i \stackrel{ind}{\sim} N_k(\mathbf{m}_i, \mathbf{S}_i), \quad i = 1, \dots, n, \quad (3)$$

where, $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n$ are an i.i.d. sample of latent variables from the mixing distribution G ,

$$\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n|G \stackrel{iid}{\sim} G. \quad (4)$$

The model is completed with a prior probability model for the random distribution G . We assume

$$G|\alpha, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \mathbf{D} \sim \mathcal{D}(\alpha G_0), \quad (5)$$

a DP prior with total mass parameter α and baseline distribution G_0 . For the baseline distribution G_0 we assume a joint distribution of a k -dimensional normal and an independent inverse Wishart. Specifically, we take $G_0(\mathbf{m}, \mathbf{S}) = N_k(\mathbf{m}|\boldsymbol{\lambda}, \boldsymbol{\Sigma}) \text{IWish}_k(\mathbf{S}|\nu, \mathbf{D})$, where $N_k(\mathbf{x}|\dots)$ and $\text{IWish}_k(\mathbf{A}|\dots)$ indicate, respectively, a k -dimensional normal distribution for the random vector \mathbf{x} and an inverse Wishart distribution for the $k \times k$ (symmetric and positive definite) random matrix \mathbf{A} . One of the attractive features of the DP prior is that

it allows straightforward posterior inference with MCMC simulation. The computational effort is, in principle, independent of the dimensionality of $\boldsymbol{\theta}_i$. Because of its computational simplicity, the DP is by far the most commonly used prior probability model for random probability measures. The DP generates almost surely discrete measures (e.g., Blackwell and MacQueen, 1973, Sethuraman, 1994). In some applications this discrete nature of the DP is awkward. However, in our setting, the discreteness is an asset as it simplifies interpretation. Let δ_x denote a point mass at x . The DP generates a discrete measure

$$G = \sum_{h=1}^{\infty} w_h \delta_{\theta_h} \quad (6)$$

with stochastically ordered weights w_h . See Sethuraman (1994) for details. *A priori*, the first few weights cover most of the probability mass. In other words, we can calibrate the DP model to assume *a priori* only few different polychoric correlations. *A posteriori*, weights are adjusted as required by the data, adding more dispersion to the distribution of the imputed polychoric correlations as the values of the ordinal factors vary across the contingency table.

To complete the model specification, we assume independent hyperpriors

$$\alpha \sim \text{Gamma}(a_0, b_0), \quad \boldsymbol{\lambda} \sim N_k(\mathbf{q}, \mathbf{Q}), \quad \boldsymbol{\Sigma} \sim \text{IWish}_k(b, \mathbf{B}), \quad \mathbf{D} \sim \text{Wish}_k(c, \mathbf{C}), \quad (7)$$

where $\text{Gamma}(\dots)$ and $\text{Wish}_k(\dots)$ denote a Gamma and a Wishart distribution, respectively, with fixed scalar hyperparameters ν , a_0 , b_0 , b , c , a k -dimensional vector \mathbf{q} , and $k \times k$ positive definite matrices \mathbf{Q} , \mathbf{B} and \mathbf{C} .

The model is stated in terms of covariance matrices instead of correlation matrices. The polychoric correlations are induced by the covariance matrix \mathbf{S}_i . A critical advantage of using covariance matrices is to avoid the difficulties associated with the modeling of correlation matrices. See, for example, Chib and Greenberg (1998), Daniels and Kass (1999) or McCulloch et al. (2000). Also, the variances on the diagonal of \mathbf{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 \mathbf{m}_i .

Most importantly, with the nonparametric mixture $f(\cdot|G)$ for the latent variables we can model essentially any probability distribution for contingency tables. We give a con-

structive proof. Using (6) we can write the random mixture $f(\cdot|G)$ as a countable mixture of normal kernels, $\sum_{h=1}^{\infty} w_h p_{N_k}(\cdot|\mathbf{m}_h, \mathbf{S}_h)$. Consider any set of probabilities $\{p_{\ell_1 \dots \ell_k}\}$ for the contingency table. Let $\{\tilde{p}_{\ell_1 \dots \ell_k}\}$ be the corresponding set of probabilities under the mixture model $f(\cdot|G)$. Hence

$$\tilde{p}_{\ell_1 \dots \ell_k} = P \left(\bigcap_{j=1}^k \{ \gamma_{j, \ell_{j-1}} < Z_j \leq \gamma_{j, \ell_j} \} | G \right),$$

where $\mathbf{Z} = (Z_1, \dots, Z_k) \sim f(\cdot|G)$ and the cutoffs, for each j , have fixed (say unit spaced) values. For each cell (ℓ_1, \dots, ℓ_k) with $p_{\ell_1 \dots \ell_k} > 0$ center one term of the mixture $f(\cdot | G)$ within the rectangle defined by the corresponding cutoffs. For example, we could, for some h , set $m_{hj} = \gamma_{j, \ell_j} - 0.5$, $j = 1, \dots, k$, and choose \mathbf{S}_h such that $\sqrt{1 - \epsilon}$ of the mass of the $N_k(\mathbf{m}_h, \mathbf{S}_h)$ kernel is within the rectangle $(\gamma_{1, \ell_1 - 1}, \gamma_{1, \ell_1}) \times \dots \times (\gamma_{k, \ell_k - 1}, \gamma_{k, \ell_k})$. Finally, set the corresponding weights w_h equal to $\sqrt{1 - \epsilon} p_{\ell_1 \dots \ell_k}$ to obtain $|p_{\ell_1 \dots \ell_k} - \tilde{p}_{\ell_1 \dots \ell_k}| < \epsilon$, for all cells (ℓ_1, \dots, ℓ_k) , i.e., an arbitrarily accurate approximation of the probability distribution for the contingency table by means of the mixture model for the latent variables.

This feature of the model has two important implications. First, the argument above shows that 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. The model provides a very general and flexible modeling framework for contingency tables. In particular, it allows polychoric correlations to vary across the contingency table, a feature that can be revealing of underlying patterns in applications.

Secondly, our model also provides a simple way to deal with the cutoffs. The random number of components and the distinct locations and covariance matrices in the mixture yield the result above without the need to consider random cutoffs. Hence, in the implementation of the model, there is no loss of generality in assuming fixed cutoffs. An important advantage of this approach is that the typically complex updating mechanisms for cutoffs (for one such example, see Cowles, 1996) are not required.

2.2 Prior Specification

Practical use of model (2) – (7) requires adopting specific values for ν , a_0 , b_0 , b , c , \mathbf{q} , \mathbf{Q} , \mathbf{B} and \mathbf{C} .

The discrete nature of the DP realizations leads to a clustering structure defined by grouping together subjects with identical $\boldsymbol{\theta}_i$. Denote by n^* the number of resulting clusters. Then (see, e.g., Liu, 1996)

$$E(n^*|\alpha) = \sum_{i=1}^n \frac{\alpha}{\alpha + i - 1} \approx \alpha \log \left(\frac{\alpha + n}{\alpha} \right)$$

and

$$\text{Var}(n^*|\alpha) = \sum_{i=1}^n \frac{\alpha(i-1)}{(\alpha + i - 1)^2} \approx \alpha \left\{ \log \left(\frac{\alpha + n}{\alpha} \right) - 1 \right\}.$$

Using the fact that *a priori* $E(\alpha) = a_0/b_0$ and $\text{Var}(\alpha) = a_0/b_0^2$ and an additional approximation based on Taylor series expansions we get

$$E(n^*) \approx \frac{a_0}{b_0} \log \left(1 + \frac{nb_0}{a_0} \right)$$

$$\text{Var}(n^*) \approx \frac{a_0}{b_0} \log \left(1 + \frac{nb_0}{a_0} \right) - \frac{a_0}{b_0} + \left\{ \log \left(1 + \frac{nb_0}{a_0} \right) - \frac{nb_0}{a_0 + nb_0} \right\}^2 \frac{a_0}{b_0^2}.$$

Equating these expressions with prior judgement at the mean and variance of n^* we obtain two equations that we can numerically solve for a_0 and b_0 .

To specify the remaining hyperparameters, we proceed by considering model (2) – (7) when $\alpha \rightarrow 0^+$. The model reduces to a parametric model with a single common value of $\boldsymbol{\theta} = (\mathbf{m}, \mathbf{S})$ for all experimental units, i.e., $\mathbf{Z}_i \stackrel{iid}{\sim} N_k(\mathbf{m}, \mathbf{S})$, with prior $\boldsymbol{\theta}|\boldsymbol{\lambda}, \boldsymbol{\Sigma}, \mathbf{D} \sim G_0$ and the hyperpriors for $\boldsymbol{\lambda}$, $\boldsymbol{\Sigma}$ and \mathbf{D} given in (7). For each dimension of the contingency table, $j = 1, \dots, k$, let e_j and r_j denote the mean and range, respectively, of the cutoffs $\{\gamma_{j,1}, \dots, \gamma_{j,C_j-1}\}$. Let $h_j = (r_j/4)^2$ and $\mathbf{H} = \text{diag}(h_1, \dots, h_k)$. Matching (e_1, \dots, e_k) and \mathbf{H} with prior moments for \mathbf{m} we fix $\mathbf{q} = (e_1, \dots, e_k)$ and $(b - k - 1)^{-1}\mathbf{B} + \mathbf{Q} = 2\mathbf{H}$, where the left-hand sides of these two expressions arise as marginal prior moments for \mathbf{m} . Additionally, in the latter equality we used an extra variance inflation factor 2. Splitting $2\mathbf{H}$ equally between the two summands we set $(b - k - 1)^{-1}\mathbf{B} = \mathbf{Q} = \mathbf{H}$. We used $b = k + 2$ to fix the largest possible dispersion in the prior for $\boldsymbol{\Sigma}$ (subject to finite $E(\boldsymbol{\Sigma})$). To fix ν, c and \mathbf{C} , note that $E(\mathbf{S}) = (\nu - k - 1)^{-1}c\mathbf{C}$, and smaller values of $(\nu - k - 1)^{-1}c$ yield more

dispersed priors for \mathbf{D} . This observation can be used to specify ν and c (with $c \geq k$ that ensures a proper prior for \mathbf{D}). Finally, \mathbf{C} is specified by setting $E(\mathbf{S}) = \mathbf{H}$.

3 Computational Approach to Posterior Inference

Let $\boldsymbol{\theta}_i = (\mathbf{m}_i, \mathbf{S}_i)$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n)$, $\underline{\mathbf{Z}} = (\mathbf{Z}_1, \dots, \mathbf{Z}_n)$, and data = $\{\mathbf{V}_1, \dots, \mathbf{V}_n\}$. To explore the posterior distribution $p(\underline{\mathbf{Z}}, \boldsymbol{\theta}, \alpha, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \mathbf{D} \mid \text{data})$ we use a Gibbs sampling approach based on simulating from the appropriate full conditional distributions. These are obtained by considering the finite dimensional posterior that emerges after integrating out the random measure G (e.g., Blackwell and MacQueen, 1973),

$$p(\underline{\mathbf{Z}}, \boldsymbol{\theta}, \alpha, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \mathbf{D} \mid \text{data}) \propto \prod_{i=1}^n p(\mathbf{V}_i \mid \mathbf{Z}_i) \prod_{i=1}^n p_{N_k}(\mathbf{Z}_i \mid \boldsymbol{\theta}_i) p(\boldsymbol{\theta} \mid \alpha, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \mathbf{D}) p(\alpha) p(\boldsymbol{\lambda}) p(\boldsymbol{\Sigma}) p(\mathbf{D}), \quad (8)$$

where $p(\boldsymbol{\theta} \mid \alpha, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \mathbf{D})$ arises by exploiting the Polya urn characterization of the DP (Blackwell and MacQueen, 1973) and the other factors are defined by (2), (3) and (7). We provide details on some of the resulting conditional distributions and the implementation of the Gibbs sampler.

To sample latent variables $\underline{\mathbf{Z}}$, note that the full conditional of \mathbf{Z}_i depends only on \mathbf{V}_i , \mathbf{m}_i and \mathbf{S}_i and is proportional to

$$\exp \left\{ -(\mathbf{Z}_i - \mathbf{m}_i)^T \mathbf{S}_i^{-1} (\mathbf{Z}_i - \mathbf{m}_i) / 2 \right\} \prod_{j=1}^k I\{\gamma_{j, \ell_j - 1} < Z_{ij} \leq \gamma_{j, \ell_j}\},$$

where (ℓ_1, \dots, ℓ_k) is defined by the value of \mathbf{V}_i . To update \mathbf{Z}_i we draw from each of its coordinates, conditional on the rest. These conditional distributions are obtained by considering first the univariate normal that results after the conditioning is done in the non-truncated multivariate version, and truncating this to the interval $(\gamma_{j, \ell_j - 1}, \gamma_{j, \ell_j}]$.

Updating the latent mixture parameters $\boldsymbol{\theta}$ and the hyperparameters $\boldsymbol{\lambda}$, $\boldsymbol{\Sigma}$, \mathbf{D} and α proceeds with standard posterior simulation methods for DP mixtures. See, for example, MacEachern and Müller (1998).

For the following discussion it is convenient to reparameterize the latent variables $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n)$. The discrete nature of the DP implies positive probabilities for ties among the

$\theta_i = (\mathbf{m}_i, \mathbf{S}_i)$. Let $n^* \leq n$ be the number of unique values among the θ_i . Denote the set of unique values (clusters) by $\underline{\theta}^* = (\theta_1^*, \dots, \theta_{n^*}^*)$, where $\theta_r^* = (\mathbf{m}_r^*, \mathbf{S}_r^*)$. Let $\mathbf{w} = (w_1, \dots, w_n)$ be a vector of configuration indicators with $w_i = r$ if and only if $\theta_i = \theta_r^*$, and let n_r be the size of the r th cluster. Then $(\underline{\theta}^*, \mathbf{w})$ is an equivalent representation of $\underline{\theta}$, with $\theta_i = \theta_{w_i}^*$.

Implementing the earlier described MCMC algorithm we obtain posterior samples $\underline{\theta}_l^*$, \mathbf{w}_l , n_l^* , $\underline{\mathbf{Z}}_l$, α_l , λ_l , Σ_l , \mathbf{D}_l , $l = 1, \dots, L$. Straightforwardly, posterior draws n_l^* and $(\mathbf{m}_{rl}^*, \mathbf{S}_{rl}^*)$, $r = 1, \dots, n_l^*$, indicate the number of clusters, and their associated locations and polychoric correlations, suggested by the data. In fact, posteriors of polychoric correlation coefficients corresponding to a latent vector \mathbf{Z}_i are immediate from the posterior of \mathbf{S}_i , $i = 1, \dots, n$. The collection of these posteriors, for all i , enables full inference for polychoric correlations and, in particular, can be used to identify patterns in the contingency table in terms of varying values of polychoric correlations in different parts of the table. See the examples in section 4 for an illustration.

We next turn to the posterior predictive distribution for a future observation \mathbf{V}_0 . Denote by \mathbf{Z}_0 the associated latent vector. The assumptions of model (2) – (7) yield $p(\mathbf{V}_0, \mathbf{Z}_0 \mid \text{data}) = p(\mathbf{V}_0 \mid \mathbf{Z}_0) p(\mathbf{Z}_0 \mid \text{data})$, where $p(\mathbf{Z}_0 \mid \text{data})$ is the posterior predictive distribution of \mathbf{Z}_0 that can be developed using the structure induced by the DP prior. In particular,

$$p(\mathbf{Z}_0 \mid \text{data}) = \int \int p_{N_k}(\mathbf{Z}_0 \mid \mathbf{m}_0, \mathbf{S}_0) p(\mathbf{m}_0, \mathbf{S}_0 \mid \underline{\theta}^*, \mathbf{w}, \alpha, \lambda, \Sigma, \mathbf{D}) p(\underline{\theta}^*, \mathbf{w}, \alpha, \lambda, \Sigma, \mathbf{D} \mid \text{data}) \quad (9)$$

where

$$p(\mathbf{m}_0, \mathbf{S}_0 \mid \underline{\theta}^*, \mathbf{w}, \alpha, \lambda, \Sigma, \mathbf{D}) = \frac{\alpha}{\alpha + n} G_0(\mathbf{m}_0, \mathbf{S}_0) + \frac{1}{\alpha + n} \sum_{r=1}^{n^*} n_r \delta_{(\mathbf{m}_r^*, \mathbf{S}_r^*)}(\mathbf{m}_0, \mathbf{S}_0). \quad (10)$$

Having sampled from (8) and hence from $p(\underline{\theta}^*, \mathbf{w}, \alpha, \lambda, \Sigma, \mathbf{D} \mid \text{data})$, expressions (9) and (10) readily provide draws from $p(\mathbf{Z}_0 \mid \text{data})$ and Monte Carlo approximations to $p(\mathbf{z}_0 \mid \text{data})$ for any grid of values \mathbf{z}_0 . Moreover, they clarify structure and the nature and amount of learning implied by the model. Note that $p(\mathbf{Z}_0 \mid \text{data})$ emerges by averaging, with respect

to the posterior $p(\underline{\boldsymbol{\theta}}^*, \mathbf{w}, \alpha, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \mathbf{D} \mid \text{data})$, the distribution

$$p(\mathbf{Z}_0 \mid \underline{\boldsymbol{\theta}}^*, \mathbf{w}, \alpha, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \mathbf{D}) = \frac{\alpha}{\alpha + n} \int p_{N_k}(\mathbf{Z}_0 \mid \mathbf{m}, \mathbf{S}) dG_0(\mathbf{m}, \mathbf{S}) \\ + \frac{1}{\alpha + n} \sum_{r=1}^{n^*} n_r p_{N_k}(\mathbf{Z}_0 \mid \mathbf{m}_r^*, \mathbf{S}_r^*).$$

This is a mixture of multivariate normal distributions, specified by the distinct locations, \mathbf{m}_r^* , and covariance matrices \mathbf{S}_r^* (that yield the distinct sets of polychoric correlations), with an additional term that allows for new structure. The weight for this additional term decreases with increasing sample size, arguably an appealing feature of the model. As we observe more and more data, the chance of new patterns emerging in future observations decreases.

Finally, of interest is also inference for the table cell probabilities. Note that, for any cell (ℓ_1, \dots, ℓ_k) ,

$$P(V_1 = \ell_1, \dots, V_k = \ell_k \mid G) = P\left(\bigcap_{j=1}^k \{\gamma_{j, \ell_j - 1} < Z_j \leq \gamma_{j, \ell_j}\} \mid G\right) \\ = \int P\left(\bigcap_{j=1}^k \{\gamma_{j, \ell_j - 1} < Z_j \leq \gamma_{j, \ell_j}\} \mid \mathbf{Z} \sim N_k(\mathbf{m}, \mathbf{S})\right) dG(\mathbf{m}, \mathbf{S})$$

i.e., $P(V_1 = \ell_1, \dots, V_k = \ell_k \mid G)$ is a linear functional of $f(\cdot \mid G)$. Its posterior can be obtained using the approach of Gelfand and Kottas (2002). We omit details here simply noting that the approach uses posterior draws from $p(\underline{\boldsymbol{\theta}}^*, \mathbf{w}, \alpha, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \mathbf{D} \mid \text{data})$ and involves approximation of DP realizations, using the constructive definition in Sethuraman (1994), and evaluation of k -dimensional normal probabilities.

4 Data Illustrations

We present results from the analysis of two data sets in sections 4.1 and 4.2, after discussing below a simple diagnostic that can indicate when a single multivariate normal might be inadequate to model the underlying latent variables and hence suggest the nonparametric model as a better choice.

To define such a diagnostic, we can examine the likelihood that results from (2) and (3) only, ignoring the prior structure. Consider first the case $k = 2$. As in section 2.2, assume we let $\alpha \rightarrow 0^+$ so that there is only one value of $\boldsymbol{\theta} = (\boldsymbol{m}, \boldsymbol{S})$ that is common to all experimental units. Using any available program for statistical inference in contingency tables, we can estimate the polychoric correlation between latent variables. Note that this may require assuming $\boldsymbol{m} = 0$ and that \boldsymbol{S} is actually a correlation matrix, as some programs produce the MLE of cutoffs as well. The results of the estimation can be used to compute expected cell frequencies which can be compared to the observed frequencies. For example, Pearson's χ^2 statistic $\sum_c (O_c - E_c)^2 E_c^{-1}$ is a convenient simple choice, where O_c and E_c represent the observed and expected frequencies for cell c , respectively, and the summation is over all cells in the contingency table. Large values of this statistic can be interpreted as a poor fit of the one-component mixture model and therefore support larger mixtures. Alternatively, a plot of pairs (O_c, E_c) can be used for comparison. The case $k > 2$ can be handled in a similar way. However, because some computer programs can only provide pairwise polychoric correlations, it could be necessary to repeatedly run the program to form the complete matrix. After this is done, we can measure the dissimilarity between observed and expected cell frequencies as earlier.

4.1 A Simulated Data Set

We test the performance of the model using simulated data from a non-standard distribution for the underlying latent variables. Specifically, setting $k = 2$, we generated $n = 100$ latent observations from a mixture (with equal weights) of two bivariate normals with means $(-1.5, -1.5)$ and $(0.5, 0.5)$, variances $(0.25, 0.25)$ for both components, and covariances -0.175 and 0.0875 , respectively. The latent data are plotted in Figure 2. Using cutoffs $-2.5, -1.5, -0.5, 0.5, 1.5$ for both latent variables, a contingency table is generated by grouping the latent data. The observed cell frequencies are included in Table 1. The corresponding Pearson statistic for these data is 138.79, which compared to the nominal 95% quantile of the χ^2 distribution with $35 - 11 = 24$ degrees of freedom (5 cutoffs for each latent variable plus one correlation adds up to 11 parameters) provides evidence against a

single multivariate probit model.

We used the MCMC algorithm of section 3 to fit the model. Convergence, assessed using multiple chains, appears to be rather fast. In addition, posterior results are quite robust to different values of prior hyperparameters. For some illustrations, Figures 1 and 2 provide posterior results under two priors for α , a Gamma(2,0.9) distribution and a, considerably more dispersed, Gamma(2,0.1) distribution. For the other hyperparameters, following section 2.2 and based on $e_j = 0$ and $r_j = 10$, $j = 1, 2$, we take $\mathbf{q} = (0, 0)^T$, $\mathbf{H} = \text{diag}(6.25, 6.25)$, $\mathbf{Q} = \mathbf{B} = \mathbf{H}$ and $b = 4$ for the priors for $\boldsymbol{\lambda}$ and $\boldsymbol{\Sigma}$. Moreover, we set $\nu = 10$, $c = 5$ and $\mathbf{C} = \text{diag}(8.75, 8.75)$ yielding a quite dispersed prior for \mathbf{D} . The posteriors of $\boldsymbol{\lambda}$, $\boldsymbol{\Sigma}$ and \mathbf{D} (not shown), being very concentrated compared with their priors, indicate that the prior choices are vague compared to the likelihood, as well as that, at least with this sample size, the data enable learning for the hyperparameters of the model.

The model successfully captures the clustering in the cells of the table driven by the bimodal underlying distribution for latent variables. Clustering in terms of locations is depicted in Figure 2 with the Monte Carlo estimates of $p(\mathbf{z}_0|\text{data})$, evaluated using (9) and (10) and a 50×50 grid of \mathbf{z}_0 values. Clustering with regard to polychoric correlations is shown in Figure 3 where, under the Gamma(2,0.9) prior for α , we plot the posterior means of Z_{i1} against the posterior means of Z_{i2} , with the pairs assigned to each one of the panels according to the value of the posterior mean of $\rho_i = \mathbf{S}_{i,12}/(\mathbf{S}_{i,11}\mathbf{S}_{i,22})^{1/2} = \text{Corr}(Z_{i1}, Z_{i2})$, $i = 1, \dots, 100$. The ranges of values of posterior means of polychoric correlations for the two clusters are -0.712 to -0.686 , with 52 associated pairs in the left panel of Figure 3, and 0.082 to 0.177 , with the remaining 48 pairs in the right panel of Figure 3. Moreover, the posterior for n^* (Figures 1(b), 1(d)) clearly indicates the need for at least two components in the mixture model. Finally, Table 1, again under the Gamma(2,0.9) prior for α , provides means and (equal tail) 95% interval estimates for posteriors of $P(V_1 = \ell_1, V_2 = \ell_2|G)$, $\ell_1, \ell_2 = 1, \dots, 6$.

4.2 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: 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 2. The Pearson statistic for a multivariate probit model is 49.51, much beyond the 95% quantile of the χ^2 distribution with $24 - 9 = 13$ degrees of freedom. Again, we interpret these numbers as supporting our model.

To fit the model to these data, we use cutoffs $-1, 0, 1, 2$ for both variables. Again, following the approach of section 2.2, we consider $e_j = 0$ and $r_j = 10$, $j = 1, 2$, a noninformative choice given the values of cutoffs. Hence the priors for $\boldsymbol{\lambda}$, $\boldsymbol{\Sigma}$ and \boldsymbol{D} are the same as in section 4.1. In addition, we take a Gamma(2,0.9) prior for α . As expected, based on the results with the simulated data of section 4.1 and the larger sample size available here, experimentation with other prior choices, corresponding to larger r_j and more dispersed Gamma priors for α , revealed robustness of posterior results.

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 Figures 4 and 5. Figure 4(a) contains 2500 draws from the posterior predictive distribution $p(\mathbf{Z}_0|\text{data})$ and Figure 4(b) shows one draw from the posterior of n^* , $(\mathbf{m}_r^*, \mathbf{S}_r^*)$, $r = 1, \dots, n^*$. Here $n^* = 4$, the ellipsoids are determined by $(\mathbf{m}_r^*, \mathbf{S}_r^*)$ and the associated weights n_r/n for the clusters are included in the plot. The posterior of n^* is given in Figure 6(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 4. Figure 5 is analogous to Figure 3 for the simulated data, capturing here four different ranges of values for the polychoric correlation. The approximate number of pairs in each panel of Figure 5 is easier to identify from Figure 6(a) where we plot pos-

terior means and the .025 and .975 posterior percentiles for ρ_i , $i = 1, \dots, 885$. The analogous plots for $\mathbf{S}_{i,11}$ and $\mathbf{S}_{i,22}$, $i = 1, \dots, 885$, are provided in Figures 6(b) and 6(c), respectively. These plots also capture the variability in the corresponding posteriors. Finally, Figure 7 includes four of the posteriors $P(V_1 = \ell_1, V_2 = \ell_2 | G)$, for $(\ell_1, \ell_2) = (1,1)$, $(1,2)$, $(2,1)$ and $(2,2)$. Posterior summaries (means and 95% interval estimates) for all (ℓ_1, ℓ_2) are given in Table 2.

All the results indicate the utility of mixture modeling for this data set. Even though 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.

5 Discussion

We have proposed a nonparametric Bayesian approach to model multivariate ordinal data. We have introduced a DP mixture model for latent variables defining classification probabilities in the corresponding contingency table. Two features of the model were extensively discussed. 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 two examples illustrate these points.

The model can be extended in several ways. For example, a regression on covariates as in Johnson and Albert (1999) or Chen and Dey (2000) can be added to the mixture model studied here. For instance, a (linear) regression could be proposed for the mean of the latent vectors \mathbf{Z}_i reserving the nonparametric specification for the corresponding covariance matrices. Alternatively, the DP mixture specification could also include part of the regression coefficient vector (e.g., the intercept terms). These and other extensions are the subject of current work.

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, S. Ghosh and B.K. Mallick, pp. 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, S. Ghosh and B.K. Mallick, pp. 113-131, New York: Marcel Dekker.
- Chib, S., and Greenberg, E. (1998), "Analysis of Multivariate Probit Models," *Biometrika*, 85, 347-361.
- Cowles, M.K. (1996), "Accelerating Monte Carlo Markov Chain Convergence for Cumulative-link Generalized Linear Models," *Statistics and Computing*, 6, 101-111.
- 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.
- Gelfand, A.E., and Kottas, A. (2002), "A Computational Approach for Full Nonparametric Bayesian Inference under Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 11, 289-305.
- 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.
- Johnson, V.E., and Albert, J.H. (1999), *Ordinal Data Modeling*, New York: Springer.
- Liu, J.S. (1996), "Nonparametric Hierarchical Bayes via Sequential Imputations," *The Annals of Statistics*, 24, 911-930.
- MacEachern, S.N., and Müller, P. (1998), "Estimating Mixture of Dirichlet Process Models," *Journal of Computational and Graphical Statistics*, 7, 223-238.
- 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, pp. 323-338, New York: John Wiley and Sons.
- Newton, M.A., Czado, C., and Chappell, R. (1995), "Bayesian Inference for Semiparametric Binary Regression," *Journal of the American Statistical Association*, 91, 132-141.

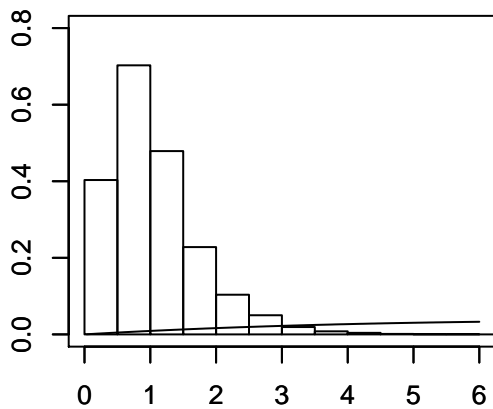
- Olsson, U. (1979), "Maximum Likelihood Estimation of the Polychoric Correlation Coefficient," *Psychometrika*, 44, 443-460.
- Read, T.R.C., and Cressie, N.A.C. (1988), *Goodness-of-fit Statistics for Discrete Multivariate Data*, New York: Springer.
- Ronning, G., and Kukuk, M. (1996), "Efficient Estimation 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.

Table 1: For the simulated data, observed cell frequencies (in bold) and posterior summaries for table cell probabilities. Rows correspond to V_1 (latent Z_1) and columns to V_2 (latent Z_2).

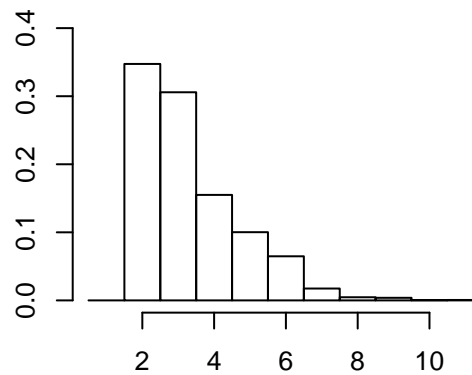
.0 .0003 (.0, .0027)	.0 .0005 (.0, .0033)	.0 .0081 (.0006, .0262)	.01 .0053 (.0005, .0179)	.0 .0004 (.0, .0027)	.0 .0001 (.0, .0011)
.0 .0008 (.0, .0051)	.05 .0680 (.0312, .1133)	.2 0.1659 (.1094, .2345)	.0 .0103 (.0013, .0295)	.0 .0003 (.0, .0022)	.0 .0002 (.0, .0014)
.01 .0140 (.0026, .0372)	.2 .1767 (.1152, .2493)	.04 .0557 (.0227, .0993)	.01 .0090 (.0008, .0261)	.0 .0049 (.0002, .0176)	.0 .0005 (.0, .0035)
.01 .0065 (.0007, .0209)	.0 .0072 (.0004, .0225)	.0 .0048 (.0002, .0174)	.14 .1169 (.0667, .1771)	.09 .1016 (.0561, .1550)	.0 .0039 (.0001, .0164)
.0 .0003 (.0, .0025)	.0 .0002 (.0, .0018)	.0 .0023 (.0, .0108)	.07 .0796 (.0399, .1312)	.15 .1313 (.0793, .1986)	.01 .0087 (.0009, .0279)
.0 .0001 (.0, .0007)	.0 .0001 (.0, .0011)	.0 .0003 (.0, .0024)	.0 .0039 (.0002, .0155)	.01 .0091 (.0013, .0258)	.0 .0019 (.0, .0118)

Table 2: For the interrater agreement data, observed cell frequencies (in bold) and posterior summaries for table cell probabilities. Rows correspond to rater A and columns to rater B.

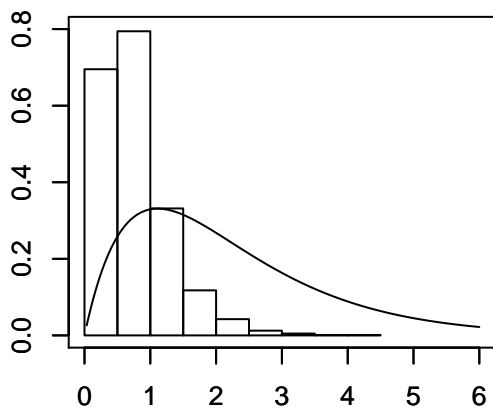
	1	2	3	4	5
1	.3288 .3261 (.2945, .3590)	.0836 .0869 (.0696, .1078)	.0011 .0013 (.0001, .0044)	.0011 .0020 (.0003, .0056)	.0011 .0007 (.0, .0027)
2	.2102 .2135 (.1858, .2428)	.2893 .2826 (.2521, .3141)	.0079 0.0082 (.0031, .0152)	.0079 .0069 (.0022, .0143)	.0034 .0031 (.0007, .0075)
3	.0023 .0023 (.0004, .0065)	.0045 .0055 (.0017, .0107)	.0 .0016 (.0003, .0038)	.0023 .0022 (.0004, .0062)	.0 .0008 (.0, .0032)
4	.0034 .0042 (.0012, .0094)	.0113 .0102 (.0042, .0187)	.0011 .0023 (.0004, .0060)	.0158 .0143 (.0066, .0240)	.0023 .0028 (.0006, .0069)
5	.0011 .0012 (.0001, .0041)	.0079 .0071 (.0026, .0140)	.0011 .0019 (.0003, .0054)	.0090 .0083 (.0034, .0153)	.0034 .0039 (.0009, .0090)



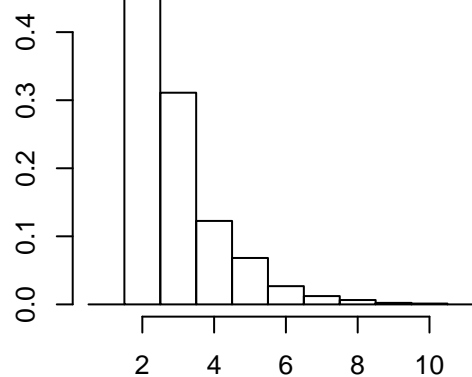
(a)



(b)



(c)



(d)

Figure 1: For the simulated data, posterior results for α and n^* under a $\text{Gamma}(2, 0.1)$ and a $\text{Gamma}(2, 0.9)$ prior for α (upper and lower panels, respectively). Priors (solid lines) and posteriors for α (panels (a) and (c)) and posteriors for n^* (panels (b) and (d)).

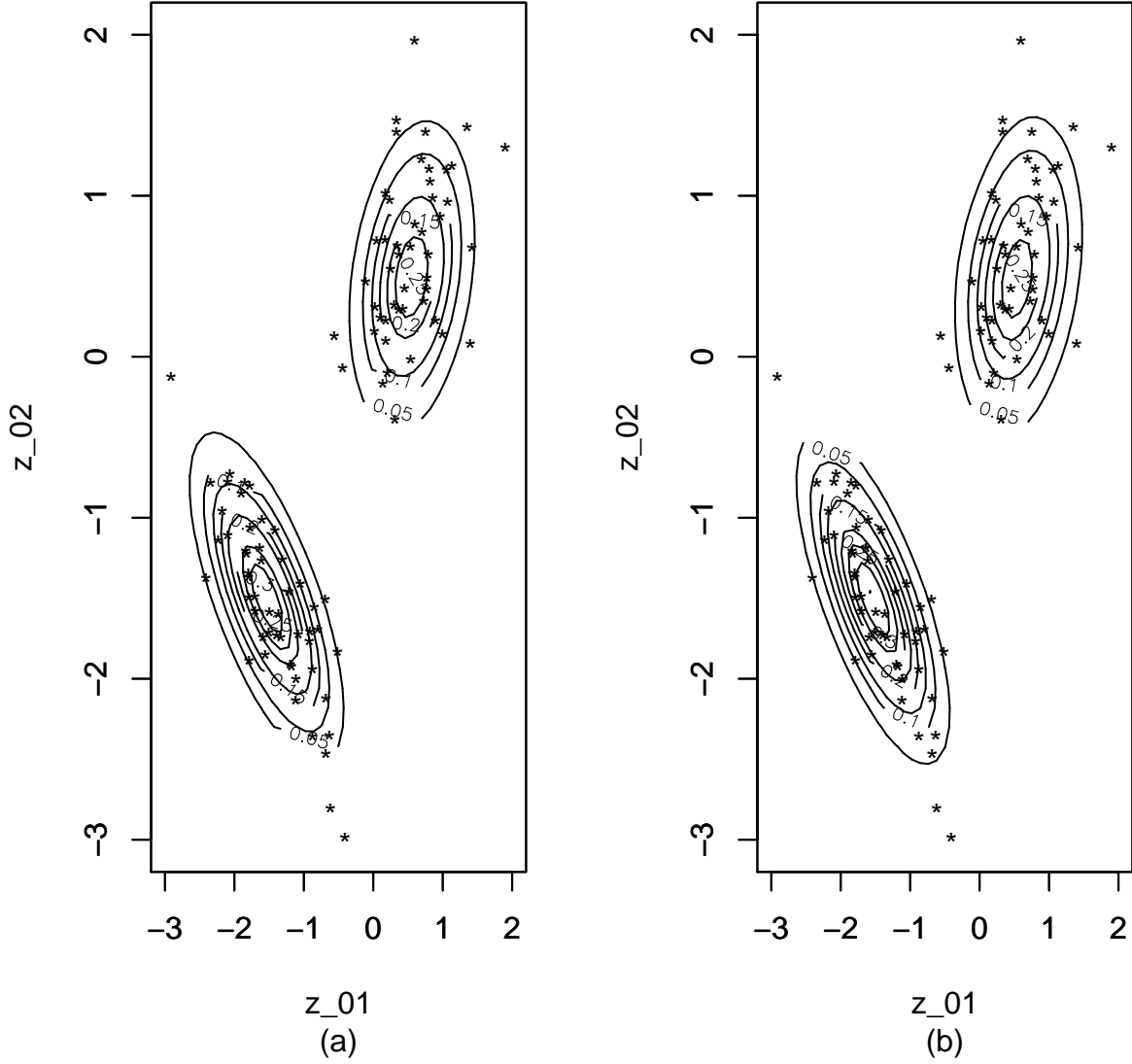


Figure 2: For the simulated data, posterior predictive density $p(\mathbf{z}_0|\text{data})$ overlaid on plot of generated latent data. Panels (a) and (b) correspond to the $\text{Gamma}(2,0.1)$ and the $\text{Gamma}(2,0.9)$ prior for α , respectively.

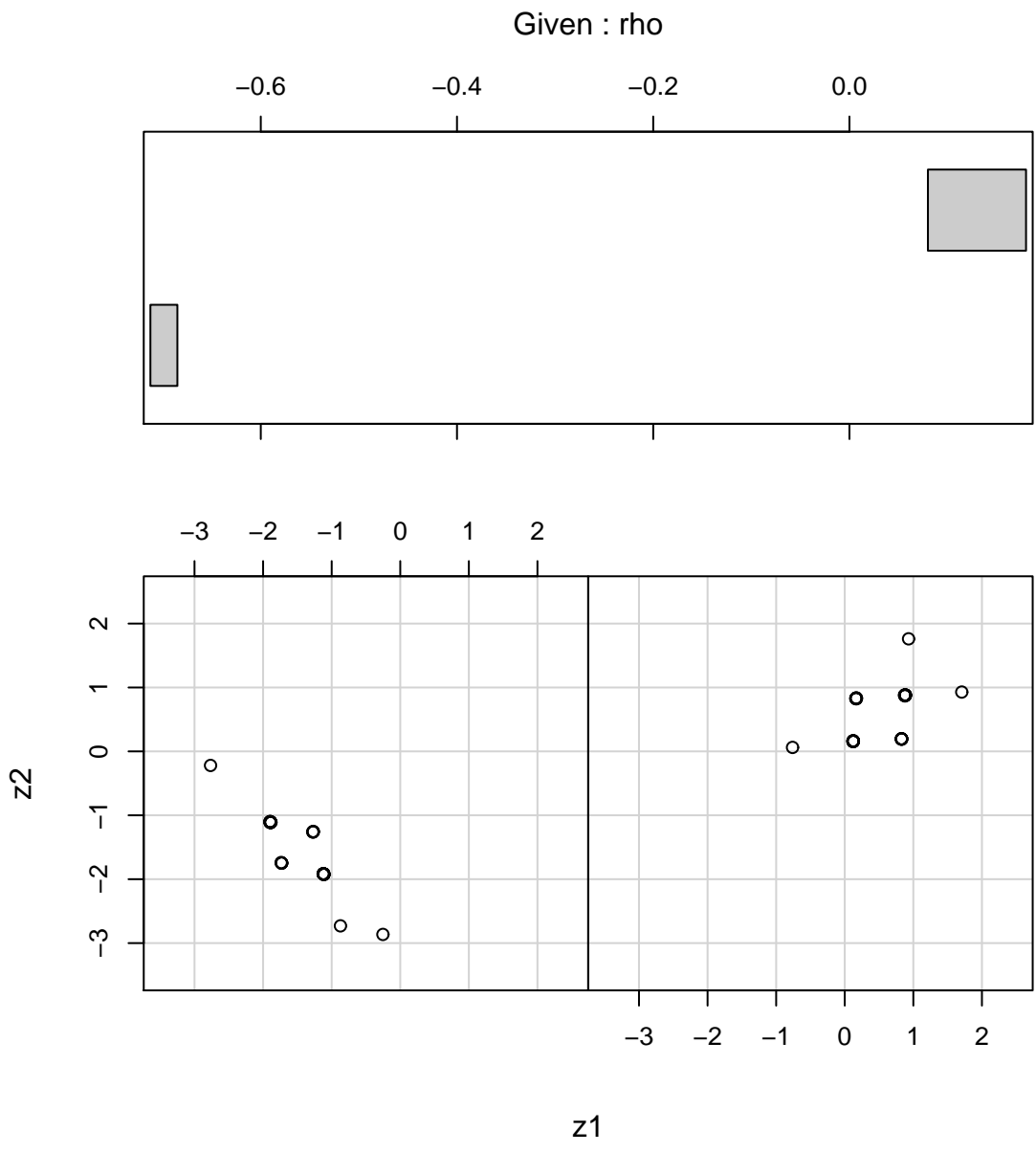


Figure 3: For the simulated data, plots of pairs of posterior means for Z_{i1} and Z_{i2} according to ranges of values of posterior means for $\rho_i = \text{Corr}(Z_{i1}, Z_{i2})$, $i = 1, \dots, 100$.

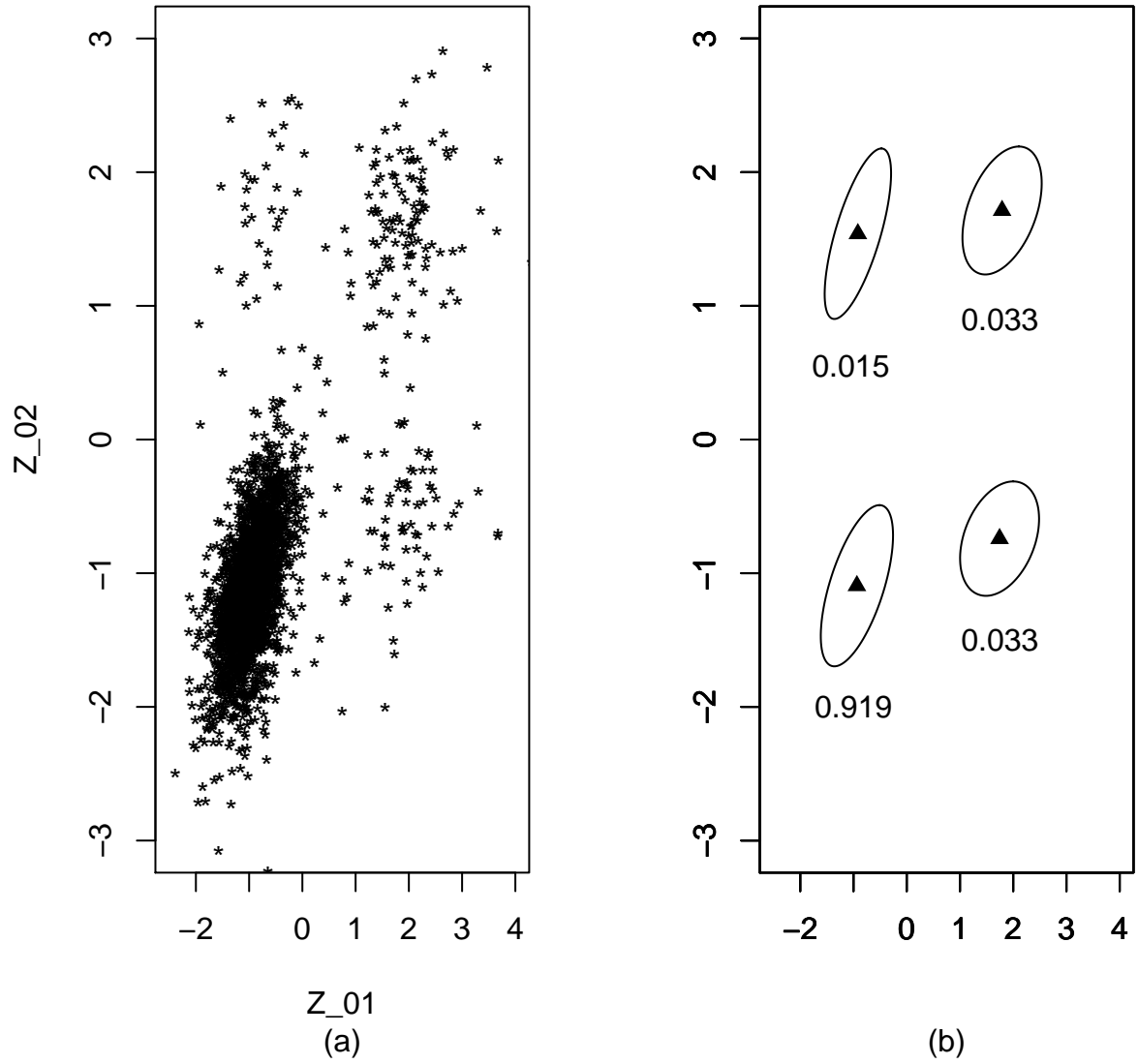


Figure 4: For the interrater agreement data, panel (a) plots draws from $p(\mathbf{Z}_0|\text{data})$. Panel (b) shows one posterior draw for $(\mathbf{m}_r^*, \mathbf{S}_r^*)$ with $n^* = 4$ clusters. The numbers below each ellipsoid are the corresponding weights n_r/n .

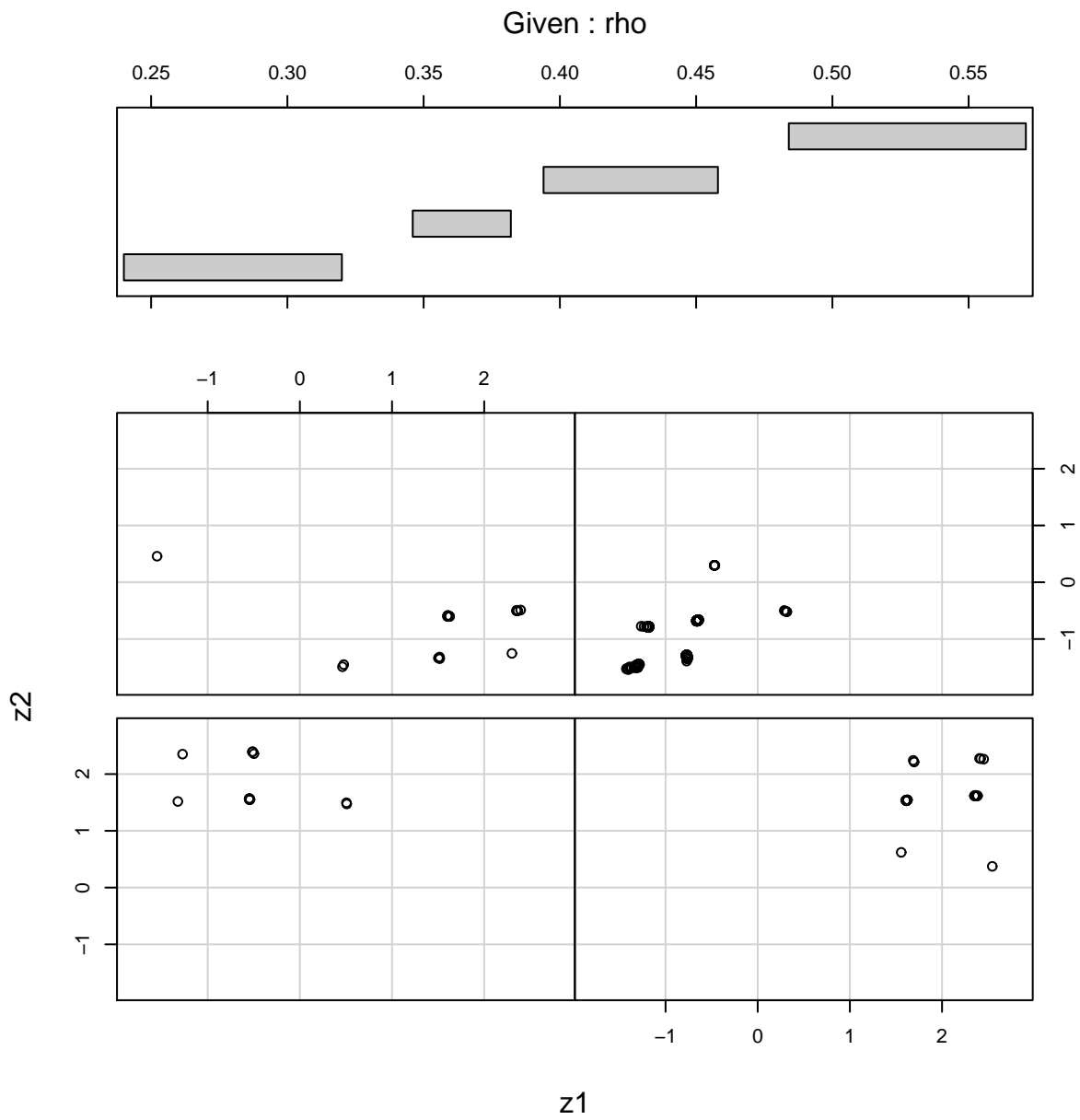


Figure 5: For the interrater agreement data, plots of pairs of posterior means for Z_{i1} and Z_{i2} for different ranges of values of posterior means for $\rho_i = \text{Corr}(Z_{i1}, Z_{i2})$, $i = 1, \dots, 885$.

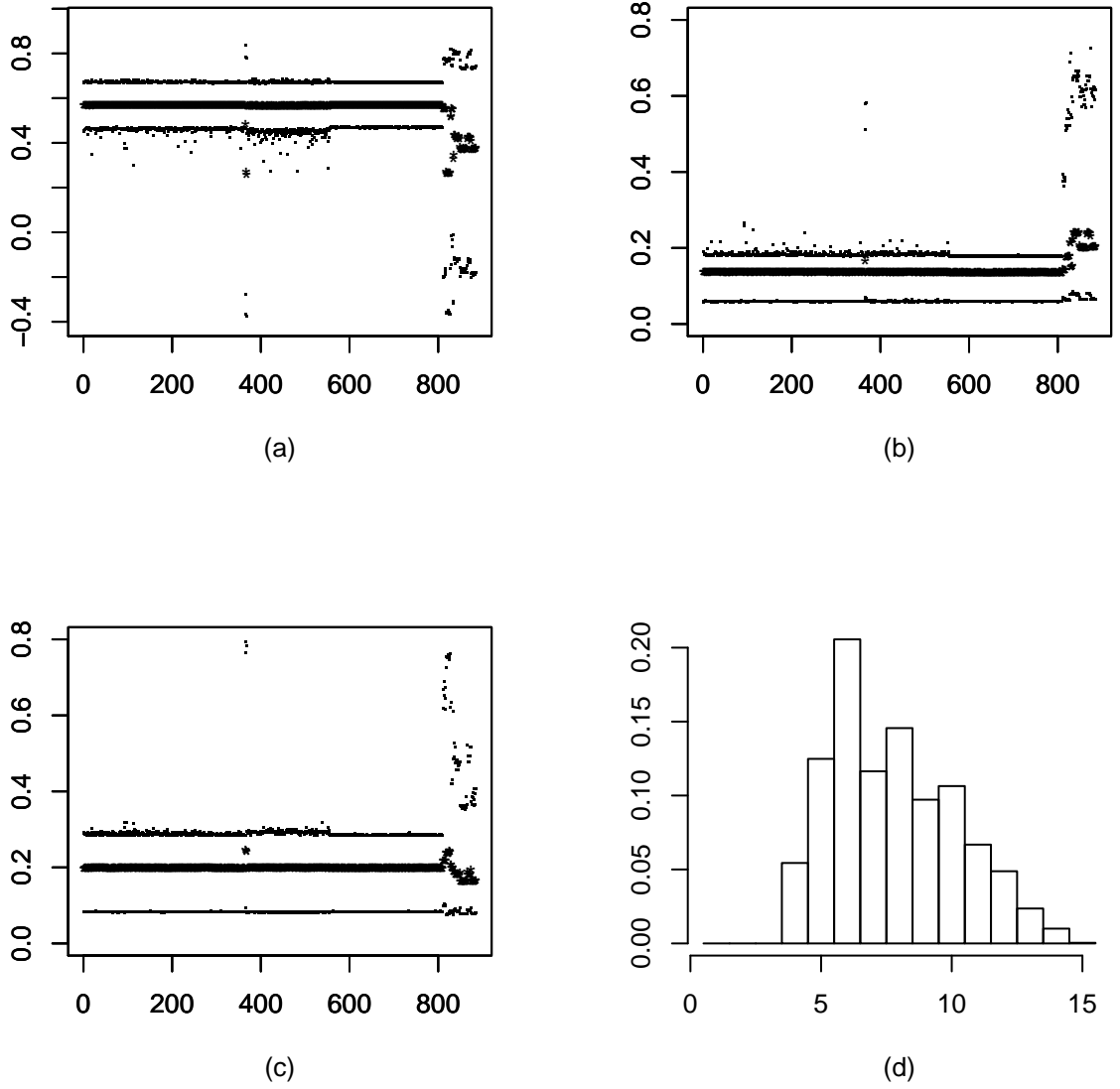


Figure 6: For the interrater agreement data, panels (a), (b) and (c) plot, for $i = 1, \dots, 885$, posterior summaries for ρ_i , $\text{Var}(Z_{i1})$ and $\text{Var}(Z_{i2})$, respectively. Panel (d) provides the posterior for n^* .

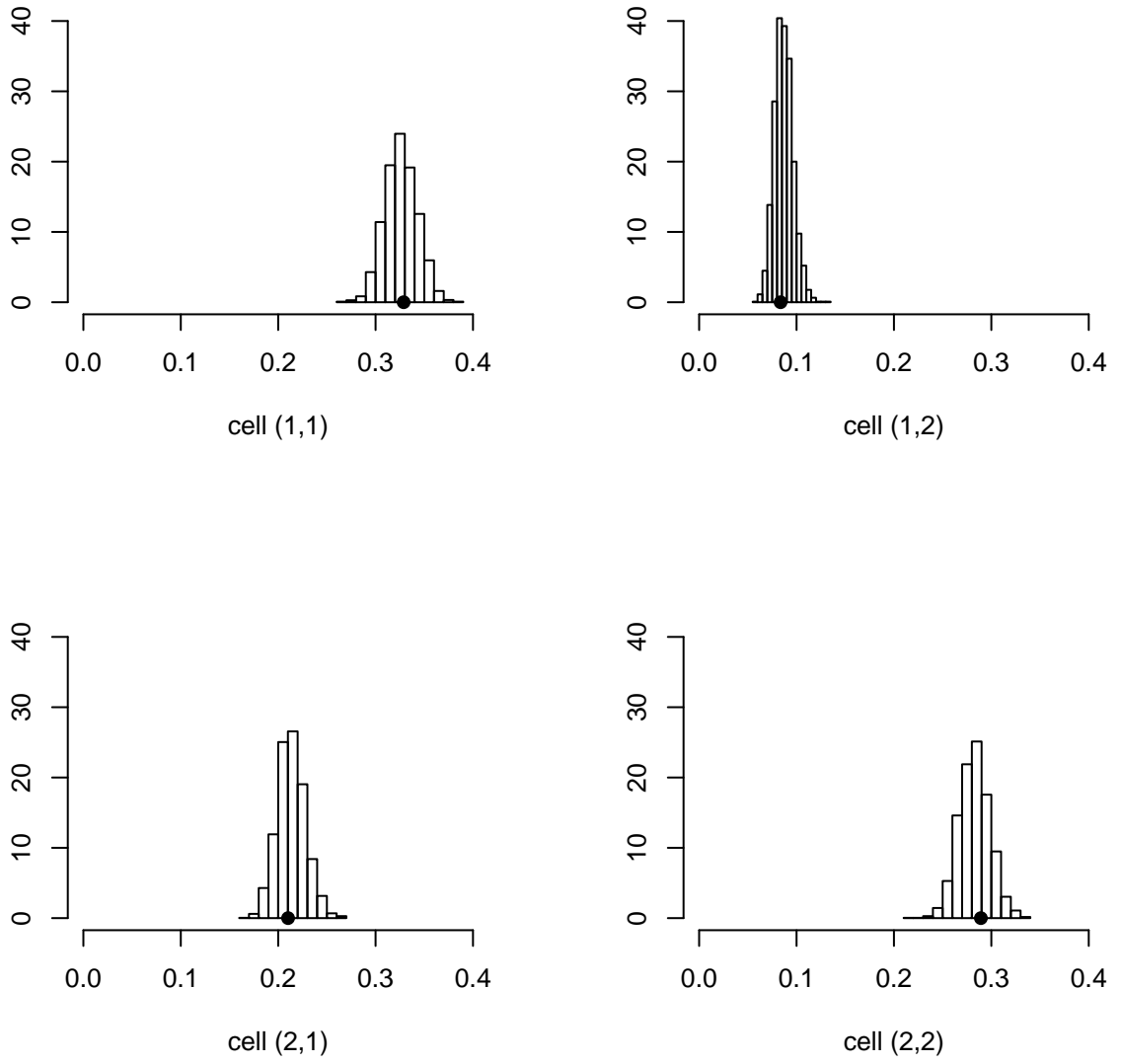


Figure 7: For the interrater agreement data, posteriors for four table cell probabilities. The circles denote the corresponding observed cell frequencies.