Tarantola, Claudia; Consonni, Guido; Dallaportas, Petros

Working Paper

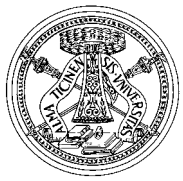
Bayesian clustering for row effects models

Quaderni di Dipartimento, EPMQ, Università degli Studi di Pavia, No. 187

Provided in Cooperation with:
University of Pavia, Department of Economics and Quantitative Methods (EPMQ)

Suggested Citation: Tarantola, Claudia; Consonni, Guido; Dallaportas, Petros (2006) : Bayesian clustering for row effects models, Quaderni di Dipartimento, EPMQ, Università degli Studi di Pavia, No. 187

This Version is available at:
http://hdl.handle.net/10419/87109



Quaderni di Dipartimento

Bayesian clustering for row effects models

Claudia Tarantola
(University of Pavia)

Guido Consonni
(University of Pavia)

Petros Dallaportas
(Athens University)

187 (11-06)

Dipartimento di economia politica
e metodi quantitativi
Università degli studi di Pavia
Via San Felice, 5
I-27100 Pavia

Novembre 2006

Bayesian clustering for row effects models

Claudia Tarantola *
University of Pavia, Italy

Guido Consonni
University of Pavia, Italy

Petros Dellaportas
Athens University of Economics and Business, Athens, Greece.

Abstract

We deal with two-way contingency tables having ordered column categories. We use a row effects model wherein each interaction term is assumed to have a multiplicative form involving a row effect parameter and a fixed column score. We propose a methodology to cluster row effects in order to simplify the interaction structure and enhancing the interpretation of the model. Our method uses a product partition model with a suitable specification of the cohesion function, so that we can carry out our analysis on a collection of models of varying dimensions using a straightforward MCMC sampler. The methodology is illustrated with reference to simulated and real data sets.

Keywords: Clustering; Contingency table; Log-linear model; Markov Chain Monte Carlo; Mixture of Dirichlet process prior; Partition; Product partition model; Row effects model.

1 Introduction

The aim of this paper is to propose a novel Bayesian methodology for the analysis of a two-way contingency table with ordered column categories. Ordered categorical variables are common in many applied areas of research, ranging from the

*Address for correspondence: Claudia Tarantola, Dipartimento di Economia Politica e Metodi Quantitativi, Via S. Felice 7, 27100 Pavia, Italy. E-mail: claudia.tarantola@unipv.it

social to the biomedical sciences, see for example Clogg and Shihadeh (1994), Etzioni *et al.* (1994) and Agresti (1999, 2002). They have been widely examined in the Bayesian literature, see for example Agresti and Chuang (1989), Evans *et al.* (1993, 1997), Brink and Smith (1996), Johnson and Albert (1999), Lang (1999), Congdon (2001) and Dellaportas and Tarantola (2005). Further references are provided in the Discussion in Section 5 of this paper.

We consider a two-way table having ordered column categories and we apply a *row effects* model, Goodman (1979). In a row effects model each interaction term is given by the product of a specific row parameter, called *row effect*, and a fixed column score. This model is appealing because it allows for an interaction term between rows and columns without becoming saturated as in the standard log-linear modelling. In this way one hopes to accommodate patterns of trend in the table, through the introduction of a set of monotone increasing column scores. On the other hand, model interpretation and parameter estimates would benefit if one could further simplify the interaction structure. In particular, as we detail in Section 2.1, it is interesting to find out which row effects can be deemed to be *equal*. In this case, the conditional distribution of the column variable given the row variable is identical for all row levels having equal effects.

We propose to achieve this objective by means of a particular *clustering* of the row effects, in a such a way that each *cluster* only contains *identical* row effects. To implement our clustering procedure, we suggest to adopt a Bayesian approach, which is model-based, produces a result which affords a probabilistic interpretation, and is highly flexible. To this end we rely on a *product partition model*, Hartigan (1990) and Barry and Hartigan (1992). Although our model allows to incorporate prior information about the clustering structure of the row effects (if this is available), we shall assume weak prior information, and let the data mostly determine the output even for moderate sample sizes.

The plan of the paper is the following: in Section 2 we describe the row effects model, establish the corresponding notation and present product partition models; in Section 3 we propose an MCMC sampling algorithm to produce a Bayesian clustering of the row effects; in Section 4 we apply our method to simulated and

real examples; finally, in Section 5, we conclude with a brief discussion.

2 Bayesian analysis of row effects models

2.1 A suitable parameterization

Consider an $a \times b$ contingency table that cross-classifies a sample of N subjects on two categorical variables A (nominal) and B (ordinal). We assume that each cell count n_{ij} is a Poisson random variable with expectation μ_{ij}, i.e. $n_{ij}|\mu_{ij} \overset{ind}{\sim} Po(\mu_{ij})$. The row effects model has the form

$$\log \mu_{ij} = \widetilde{\lambda_\emptyset} + \widetilde{\lambda}_i^A + \widetilde{\lambda}_j^B + \widetilde{\eta}_i v_j,$$

with $\{v_j, j = 1, \ldots, b\}$ fixed column scores, $v_1 \leq v_2 \leq \ldots \leq v_b$. The column scores are assigned following an integer-scoring method (Powers and Xie, 2000), that is we assume that the distance between any two adjacent categories is uniform across all possible values. The particular values assigned are inconsequential, as long as they are uniformly spaced. That is $v_j = j$ yields to the same model as $v_j = M \times j$ for any integer M. In the following, without loss of generality we set $v_j = j$. For identifiability reasons zero-sum constraints are typically imposed, i.e. $\sum_{i=1}^a \widetilde{\lambda}_i^A = 0$, $\sum_{j=1}^b \widetilde{\lambda}_j^B = 0$ and $\sum_{i=1}^a \widetilde{\eta}_i = 0$.

For our clustering problem, it is convenient to consider the following alternative parameterization

$$\log \mu_{ij} = \lambda_i^A + \lambda_j^B + \eta_i v_j, \tag{1}$$

with $\sum_{i=1}^a \lambda_i^A = 0$, $\sum_{j=1}^b \lambda_j^B = 0$, so that $\lambda_1^A = -\sum_{i=2}^a \lambda_i^A$, and similarly for λ_1^B. In this way no constraints are imposed on the $\eta-$parameters.

The parameters η_i are called *row effects* (not to be confused with the main effects for rows, λ_i^A with $i = 1, \ldots, a$), and they represent the main object of our analysis. It is therefore instructive to provide an interpretation of their role within model (1), see also Agresti (2002, sect. 9.5.2).

Recall that A is the nominal row variable and B the ordinal column variable. It is easy to check that the logit for adjacent categories of variable B takes the

form

$$\log \frac{\Pr(B = j+1|A = i)}{\Pr(B = j|A = i)} = (\lambda^B_{j+1} - \lambda^B_j) + \eta_i. \tag{2}$$

Because of the additive structure exhibited by (2), plots of these adjacent logits against the levels of A, are parallel piece-wise linear functions. For this reason Goodman (1983) referred to (1) as a *parallel odds* model. For an illustration of this feature of the model see Section 4.2. Notice that the logarithm of *odds-ratios* for adjacent categories of variable B result in differences between the corresponding row effects, e.g.

$$\eta_k - \eta_i = \log \frac{\Pr(B = j+1|A = k)}{\Pr(B = j|A = k)} - \log \frac{\Pr(B = j+1|A = i)}{\Pr(B = j|A = i)} \quad \text{with } k > i, \tag{3}$$

independently of j. In particular, when $\eta_i = \eta_k$ rows i and k have identical conditional distributions, while if $\eta_i > \eta_k$ B is stochastically larger in row i than in row k. For an interesting interpretation of the equality $\eta_i = \eta_k$ in terms of *mergings* of rows see the Discussion in Section 5 of this paper. The above remarks motivate our objective of clustering the row-effects η_i: not only will the ensuing model be more parsimonious but interpretability will be significantly enhanced, especially if the clustering is substantial.

For computational purposes, model (1) can be written in matrix notation as follows

$$\log \mu = D\theta = X\lambda + V\eta, \tag{4}$$

where D is the design matrix having full column rank, $D = [X \vdots V]$, $\theta = [\lambda', \eta']'$, $\lambda = [\lambda^A_2, \lambda^A_3 \ldots, \lambda^A_a, \lambda^B_2, \lambda^B_3 \ldots, \lambda^B_b]'$, $\eta = [\eta_1, \eta_2, \ldots, \eta_a]'$. Note that the matrices X and V have both full column rank, but only matrix X satisfies the sum-to-zero constraints.

For example, a model for a 4×2 contingency is given by

$$\begin{pmatrix} \log(\mu_{11}) \\ \log(\mu_{21}) \\ \log(\mu_{31}) \\ \log(\mu_{41}) \\ \log(\mu_{12}) \\ \log(\mu_{22}) \\ \log(\mu_{32}) \\ \log(\mu_{42}) \end{pmatrix} = \begin{pmatrix} -1 & -1 & -1 & -1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 1 \\ -1 & -1 & -1 & 1 & 2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 2 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 2 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 2 \end{pmatrix} \times \begin{pmatrix} \lambda_2^A \\ \lambda_3^A \\ \lambda_4^A \\ \lambda_2^B \\ \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{pmatrix}.$$

2.2 Product partition models

Product partition models (PPMs) are related to partition models, although the latter term seems to refer to a variety of situations, see for example Malec and Sedransk (1992), Consonni and Veronese (1995), Denison *et al.* (2002). More recent papers on PPMs include Crowley (1997), Loschi *et al* (2003), Quintana and Iglesias (2003).

We now briefly review the theory on PPMs with reference to our specific problem. Let $S_0 = \{1, \ldots, a\}$ be the set of rows of a contingency table. To each row i is associated a vector of counts $n_i = (n_{i1}, \ldots, n_{ib})'$. We indicate with $n = (n_1', \ldots, n_a')'$ the combined vector of cell counts. A partition $\rho = \{S_1, \ldots, S_{|\rho|}\}$ of the set S_0 is defined by the property that $S_d \cap S_{d'} = \emptyset$ for $d \neq d'$ and $\cup_d S_d = S_0$ (for a given finite set U we denote with $|U|$ the number of elements in U). Given a partition ρ, we assume that all η_i pertaining to the same set $S_d \in \rho$ are equal. In the following, see in particular Section 3.1, we shall use the term *cluster* to denote a collection of η_i whose value is the same since their subscripts belong to the same subset within a given ρ. The number of all possible partitions is given by $B(a)$, the Bell number of order a, recursively defined by $B(a+1) = \sum_{k=0}^{a} \binom{a}{k} B(k)$ with $B(0) = 1$.

Each partition ρ is assigned a prior probability given by

$$P(\rho = \{S_1, \ldots, S_{|\rho|}\}) = K \prod_{d=1}^{|\rho|} C(S_d), \tag{5}$$

where $C(S_d)$ is a cohesion function and K is the normalizing constant. Equation (5) is referred to as the *product distribution* for partitions.

Let $n_{S_d} = \{n_i : i \in S_d\}$ denote the vector of cell counts for all rows belonging to the same set S_d. For a given ρ, the conditional distribution of n_{S_d}, $p_{S_d}(n_{S_d}|\rho)$, is assumed to depend only on S_d and not on the other sets of the partition. Moreover, given $\rho = \{S_1, \ldots, S_{|\rho|}\}$, the counts $n_{S_1}, \ldots, n_{S_{|d|}}$ are assumed to be independent with distribution

$$p(n|\rho) = \prod_{d=1}^{|\rho|} p_{S_d}(n_{S_d}|\rho). \tag{6}$$

Equations (5) and (6) uniquely determine the joint law of (n, ρ). The corresponding posterior distribution of ρ is again of the form (5) with (posterior) cohesions $C(S_d)p_{S_d}(n_{S_d}|\rho)$.

The cohesions can be specified in different ways; here we follow the approach presented by Quintana and Iglesias (2003) and set

$$C(S_d) = c \times (|S_d| - 1)!, \tag{7}$$

with $c > 0$. This choice can be justified considering the connection between parametric PPMs and the class of Bayesian nonparametric models based on mixture of Dirichlet Processes (Antoniak, 1974). Under the latter prior, the marginal distribution of the observables is a specific PPM with the cohesion functions specified by equation (7), see Quintana and Iglesias (2003). Efficient MCMC algorithms have been developed for Bayesian nonparametric problems based on Mixtures of Dirichlet Processes, see e.g. West *et al.* (1994), Escobar and West (1995), Bush and MacEachern (1996), MacEachern and Müller (1998, 2000) and Jan and Neal (2004). The connection between parametric PPMs and nonparametric models with a Dirichlet process prior suggests to adapt these algorithms to our problem.

2.3 A hierarchical model

We consider the following model

$$
\begin{aligned}
& n_{ij}|\rho,(\phi_1,\ldots,\phi_{|\rho|}),\lambda \overset{ind}{\sim} Po(\mu_{ij}) \\
& \log\mu = X\lambda + V\eta \\
& \lambda \sim N(0,\sigma^2_\lambda I) \\
& \phi_1,\ldots,\phi_{|\rho|}|\rho,\sigma^2_\phi \overset{iid}{\sim} N(0,\sigma^2_\phi) \\
& \rho \sim \text{product distribution, with } \; C(S_i) = c\times(|S_i|-1)! \\
& \sigma^2_\phi \sim IG(c_\phi, d_\phi),
\end{aligned}
$$

where $\phi = (\phi_1,\ldots,\phi_{|\rho|})'$ is the vector of all distinct values of $\eta = (\eta_1,\ldots,\eta_a)'$ for a given partition ρ, the parameter σ^2_λ is fixed, the product distribution is defined in (5), and $IG(c_\phi, d_\phi)$ is an inverted gamma distribution with expectation $d_\phi/(c_\phi-1)$, $c_\phi > 1$ and $d_\phi > 0$.

The joint distribution of the variables involved in the model is thus given by

$$
\begin{aligned}
p(n,\lambda,\eta,\rho=\{S_1,\ldots,S_k\},\sigma^2_\phi) &\propto \left\{\prod_{i=1}^{a}\prod_{j=1}^{b}(\mu_{ij})^{n_{ij}}\right\}\exp\left\{-\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}\right\} \\
\times \quad & \exp\left\{-\frac{1}{2}\frac{\lambda'\lambda}{\sigma^2_\lambda}\right\}\times\frac{1}{(\sigma^2_\phi)^{\frac{|\rho|}{2}}}\exp\left\{-\frac{1}{2\sigma^2_\phi}\sum_{d=1}^{|\rho|}(\phi_d)^2\right\} \\
\times \quad & \left(\sigma^2_\phi\right)^{-(c_\phi+1)}\exp\left\{-\frac{d_\phi}{\sigma^2_\phi}\right\}\times\prod_{i=1}^{|\rho|}C(S_d).
\end{aligned}
$$

3 MCMC computation

We propose to sample from the joint posterior distribution of model and parameters using the following MCMC algorithm.

STEP 1. Update (ρ,η) applying the *No Gaps Algorithm* described in Section 3.1

STEP 2. Update λ using a Metropolis step. We apply a random walk Metropolis algorithm with proposal density $q(\cdot)$, a multivariate normal distribution with mean equal to the current value of λ and variance matrix equal to the (properly scaled) conditional maximum likelihood estimated covariance

matrix. The latter is computed using a profile likelihood with η set to the current value and applying a Newton-Raphson algorithm.

STEP 3. Sample from the full conditional of σ_ϕ^2 which is

$$IG\left(c_\phi + \frac{|\rho|}{2}, d_\phi + \frac{\sum_{d=1}^{|\rho|}(\eta^{S_d})^2}{2}\right).$$

3.1 "No Gaps Algorithm"

To update the partition structure ρ we apply the *No Gaps Algorithm* by MacEachern and Müller (1998, 2000). We briefly review this sampling scheme and illustrate it with reference to our specific problem.

We start out by fixing the notation. For given ρ, let $(\phi_F', \phi_E')'$ be the vector of all *clusters*, with $\phi_F = (\phi_1, \ldots, \phi_{|\rho|})'$ the set of distinct values of η_i (*full* clusters) and $\phi_E = (\phi_{|\rho|+1}, \ldots, \phi_a)'$ the set of potential but not yet used cluster locations (*empty* clusters). Let $s = (s_1, s_2, \ldots, s_a)'$ denote the vector of configuration indicators, that is $s_i = \ell$ if and only if $\eta_i = \phi_\ell$, $i = 1, \ldots, a$, $\ell = 1, \ldots, |\rho|$, and let $|S_\ell|$ be the number of all s_i that equal ℓ. We recall that with the term cluster we refer to the set of all η_i with identical configuration location. There are no gaps in the values of s_i, that is $|S_\ell| > 0$ for $\ell = 1, \ldots, |\rho|$ and $|S_\ell| = 0$ for $\ell = |\rho| + 1, \ldots, a$. In the following formulas, a vector with subscript "$-i$" means that the ith element has been removed; the superscript "$-i$" instead refers to a summary index with the ith element removed. In particular, η_{-i} is the vector η without the ith component, while $|\rho^{-i}|$ refers to the number of clusters formed by η_{-i} and $|S_\ell^{-i}|$ represents the number of elements in cluster ℓ when the element i is removed.

The algorithm is described below. For $i = 1, \ldots, a$, do Step (i) then do Step (ii).

Step (i) : If $|S_{s_i}| > 1$, then resample s_i from

$$Pr(s_i = \ell | s_{-i}, \phi, \lambda, n) \propto \begin{cases} |S_\ell^{-i}| \times f(n_i|\phi_\ell, \lambda) & \text{for } \ell = 1, \ldots, |\rho^{-i}| \\ \frac{C(\{i\})}{|\rho^{-i}|+1} \times f(n_i|\phi_\ell, \lambda) & \text{for } \ell = |\rho^{-i}| + 1. \end{cases}, \quad (8)$$

with $C(\{i\})$ defined in (7).

If $|S_{s_i}| = 1$ then with probability $(|\rho| - 1)/|\rho|$ leave s_i unchanged. Otherwise relabel clusters in such a way that $s_i = |\rho|$ and then resample s_i with probability (8). Note that if a cluster ℓ with $|S_\ell| = 1$ is removed by resampling s_i, we keep the old value of ϕ_ℓ recorded as $\phi_{|\rho|}$. The values of $(\phi_1, \ldots, \phi_a)'$ are never changed during the execution of Step (i), except for relabeling the indexes if necessary.

Step (ii) : Resample ϕ_ℓ conditionally on the configuration s and all the other parameters. For $\ell = |\rho| + 1, \ldots, a$ we simply sample from the prior $g_0(\phi_\ell|\sigma^2_\phi)$. For $\ell = 1, \ldots, |\rho|$ we sample from the conditional posterior of ϕ_ℓ

$$g(\phi_\ell|n, \lambda, \sigma^2_\phi) \propto \left[\prod_{i \in S_\ell} f(n_i|\lambda, \phi_\ell)\right] \times g_0(\phi_\ell|\sigma^2_\phi)$$

$$\propto \prod_{i \in S_\ell} \left[\left\{\prod_{j=1}^{b} \mu_{ij}^{n_{ij}}\right\} \exp\left\{-\sum_{j=1}^{b} \mu_{ij}\right\}\right] \times \exp\left\{-\frac{1}{2\sigma^2_\phi}(\phi_\ell)^2\right\}, \quad (9)$$

with $\mu_{ij} = \exp\{\lambda^A_i + \lambda^B_j + \eta_i v_j\}$.

To sample from (9) we apply a Metropolis step. The proposal distribution is a normal density with parameters based on maximum likelihood estimates of the current model calculated with a Newton-Raphson algorithm.

After drawing a new vector of locations s through Step (i), the corresponding vector η is obtained together with partition $\rho = \{S_1, \ldots, S_{|\rho|}\}$.

4 Examples

The methodology described in the previous sections is now illustrated on a simulated 10×5 table and on two real data sets. Following Quintana and Iglesias (2003) we set $c = 1$ in equation (7) in order to favour partitions with a small number of large subsets. We used a weakly informative prior on λ setting $\sigma^2_\lambda = 10\,000$, and set $c_\phi = 3$ and $d_\phi = 2$. The results obtained are rather insensitive to different choices of c_ϕ and d_ϕ.

Convergence of the MCMC algorithm was assessed using diagnostics implemented in the package BOA, see Smith (2001). In particular, the multivariate

scale reduction factor proposed in Brooks and Gelman (1998) for the three examples was equal to 1.0013, 1.0015 and 1.0521 respectively. Other diagnostic checks (not reported here) showed no specific indication of abnormal behaviour. Finally, convergence has been also checked by means of several plots of output values.

For the three examples discussed in the next sections, involving a 10×5, a 4×4 and a 12×2 contingency table, the algorithm required 20, 15 and 30 minutes respectively per 100 000 iterations on a Pentium 4 3.4 GHz personal computer. The programs were written in MATLAB; it is expected that a lower level programming language could speed up the execution time by a factor of at least 5.

4.1 Simulated data

We report here one out of the many simulated examples we experimented with. We considered model (4) with fixed parameters

$$\lambda = (2.5, -1.5,\ 1, -1.5,\ 2, -1.5,\ 1, -1,\ 1,\ 2.5, -0.5, -1.5, -3)'$$
$$\eta = (2,\ 0.5,\ 2,\ 0.5,\ 2,\ 0.5,\ 2,\ 0.5,\ 2,\ 1)',$$

calculated the vector of expected cell counts μ, and used them (suitably rounded) as cell counts (see Table 1). This was done to avoid possible confoundings due to simulation error. Notice that the true partition for the set of 10 rows is given by $\rho = \{\{1, 3, 5, 7, 9\}, \{2, 4, 6, 8\}, \{10\}\}$. We considered a run with 300 000 sweeps and a burn-in of 30 000. Table 2 presents the results for those "models", i.e. partitions, whose posterior probabilities exceed the threshold 0.01. The MCMC standard errors of the model probability estimates were calculated by splitting the Markov chain output into batches, see Geyer (1992). Notice that almost 94% of the posterior probability is concentrated on the true partition.

In Fig. 1 we report the ergodic means for the highest four posterior model probabilities (the three bottom traces can hardly be distinguished because of their very similar low values).

TABLE 1 ABOUT HERE

TABLE 2 ABOUT HERE

FIGURE 1 ABOUT HERE

Conditionally on the partition with highest posterior probability we obtained the following parameter estimates

$$\lambda = (2.51, -1.51,\ 1.01, -1.50,\ 2.01, -1.51,\ 1.01, -1.01,\ 1.00,\ 2.50, -0.50, -1.50, -3.00)'$$
$$\eta = (2.00,\ 0.49,\ 2.00,\ 0.49,\ 2.00,\ 0.49,\ 2.00,\ 0.49,\ 2.00,\ 1.00)',$$

which can be seen to be in excellent agreement with the true input values.

4.2 Premarital sex data

We consider a 4×4 table presented in Agresti (2002, p. 368). Subjects were asked their opinion about a man and woman having sexual relations before marriage (Always wrong, Almost always wrong, Wrong only sometimes, Not wrong at all). They were also asked whether methods of birth control should be available to teenagers between the age of 14 and 16. The data are the top number in each cell of Table 3 (source: General Social Survey, National Opinion research Center, Chicago, 1991).

TABLE 3 ABOUT HERE

Notice that both the row and column variable in this data set are ordinal and they play a symmetric role. For the sake of this analysis, we set the variable representing the opinion on Teenage birth control as column variable, whereas the levels of the variable on premarital sex represent the rows. Since our clustering method does not use the information contained in the ordinal nature of the row variable, we concede that more specialized models could be applied to this data set; in particular, within our setting, only the set of *contiguous* partitions of row levels should be taken into account (for instance $\{\{1\}, \{2, 3\}, \{4\}\}$ is a contiguous partition, while $(\{\{1, 3\}, \{2\}, \{4\}\}$ is not). On the other hand, our method does not forbid such partitions, and it is interesting to verify to what extent contiguous partitions might emerge empirically through our clustering procedure. The

results are encouraging in this respect: based on 100 000 sweeps with a burn-in of 10 000, Table 4 shows that the posterior distribution on the space of partitions is concentrated on two elements only, namely partition $\{\{1,2\},\{3\},\{4\}\}$ (with a probability of 91%), and the "trivial" partition $\{\{1\},\{2\},\{3\},\{4\}\}$. Notice that the prevailing partition is contiguous, since it assigns levels 1 and 2, corresponding to the opinion on premarital sex "Always wrong" and "Almost always wrong", to the same cluster. Fitted values under a standard model which assumes independence of A and B, as well as those under our model corresponding to partition $\{\{1,2\},\{3\},\{4\}\}$ are reported in Table 3: clearly our product partition model fits much better. Fig. 2 depicts the ergodic means for the two most probable models.

TABLE 4 ABOUT HERE

FIGURE 2 ABOUT HERE

To better appreciate the implications of our product partition model we report in Table 5 the posterior expectations of the model parameters together with their posterior standard deviations.

TABLE 5 ABOUT HERE

The estimated odds of Disagree instead of Strongly disagree, or Agree instead of Disagree, or Strongly agree instead of Agree, on the issue of Teenage birth control are the *same* both for people who believe that Premarital sex is Always wrong or Almost always wrong. The above odds change when considering people who believe that Premarital sex is Wrong only sometimes or Not wrong at all. Letting $\hat{\eta}_i$ denote the posterior expectation of η_i, we note that, since $(\hat{\eta}_k - \hat{\eta}_i) > 0$ for $k = 3, 4$ and $i = 1, 2$, those who believe that Premarital sex is Always wrong/Almost always wrong are also more conservative on Teenage birth control. For example, since $(\hat{\eta}_3 - \hat{\eta}_1) = 0.5125$, the estimated odds of Disagree instead of Strongly disagree, or Agree instead of Disagree, or Strongly agree instead of Agree, on the issue of Teenage birth control for people who believe that Premarital sex is Wrong only sometimes is $\exp(0.5125) = 1.67$

times the corresponding odds for people who believe that Premarital sex is Always wrong. Similarly, for those who believe that Premarital sex is Not wrong at all, the estimated odds of Disagree instead of Strongly disagree, or Agree instead of Disagree, or Strongly agree instead of Agree, on the issue of Teenage birth control, is $\exp(\hat{\eta}_4 - \hat{\eta}_1) = 2.23$ times the corresponding odds for people who believe that Premarital sex is Always wrong. Of course, since $\hat{\eta}_1 = \hat{\eta}_2$, we could have equally considered as baseline those who believe that Premarital sex is Almost always wrong instead of those who consider it Always wrong.

Fig. 3 shows the parallelism of the estimated logits for the row effects model.

FIGURE 3 ABOUT HERE

4.3 Marine corps data

We consider a data set examined by Leonard and Novick (1986) which cross-classifies 5 646 marines by School attended and Grade reported on a military aptitude test. The original table had 12 rows (School: A through L) and 8 columns (Grade: 1 highest and 8 lowest). A preliminary analysis performed by the Authors suggested that the dimension of the table might be usefully reduced by grouping together grades 1-3, corresponding to scores "above average", into a single score which we call "High", and similarly for the remaining five grades 4-8, corresponding to "below average" scores, which we denote by "Low ", so that the resulting table has dimension 12×2 (see Table 6). Furthermore, Leonard and Novick (1986) proposed a clustering of the schools in three groups $\{B, C, E, I\}$ $\{A, D, G, H\}$ and $\{F, J, K, L\}$, based on a descriptive analysis of the interaction structure between School and Grade.

TABLE 6 ABOUT HERE

The first output of our analysis is presented in Table 7. It reports those models whose posterior probability exceeds the threshold 0.01. The results are based on 500 000 iterations with a burn-in of 50 000 iterations.

TABLE 7 ABOUT HERE

The top partition, whose probability is about 58%, presents six clusters, while

the second partition, whose probability is approximately 20%, contains only five clusters. The difference between these two partitions is fairly limited: both recognize that schools $\{B, E, I\}$ form one cluster and similarly for $\{F, J\}$. They also agree on the fact that schools $\{A, G, H\}$ should belong to the same group; however school D which in the top partition belongs to the same cluster as $\{A, G, H\}$, is put in a separate cluster (together with C) in the second partition. The reader can check by himself the remaining differences between the two top partitions. Notice that none of the partitions listed in Table 7 is equal to the three-cluster structure identified by Leonard and Novick (1986): indeed all models in Table 7 contain either 5 or 6 clusters. While our method identifies a finer clustering structure than Leonard and Novick's, there is however a broad agreement between our results and theirs: indeed our top partition is a refinement of theirs; on the other hand the main discrepancy exhibited in our second most probable partition is the allocation to the same cluster of schools C and D.

The estimated odds of obtaining a Low grade instead of a High grade in school i is $\exp\{\hat{\lambda}^{Grade}_{Low} - \hat{\lambda}^{Grade}_{High} + \hat{\eta}_i\} = \exp\{2\hat{\lambda}^{Grade}_{Low} + \hat{\eta}_i\}$ (since $\hat{\lambda}^{Grade}_{High} = -\hat{\lambda}^{Grade}_{Low}$), so that schools belonging to the same cluster have constant odds since their estimated value for η is the same.

For comparative purposes, we consider for each cluster of the top partition the odds of Low instead of High grade: schools in clusters $\{B, E, I\}$ and $\{C\}$ perform better (have higher grades) than schools in cluster $\{A, D, G, H\}$, whereas the performance of schools $\{F, J\}$,$\{K\}$ and $\{L\}$ is worse than that of schools $\{A, D, G, H\}$. For example, the odds of a Low instead of High grade in cluster $\{B, E, I\}$ and $\{C\}$ are respectively $\exp\{\hat{\eta}_B - \hat{\eta}_A\} = 0.6233$ and $\exp\{\hat{\eta}_C - \hat{\eta}_A\} = 0.8028$ times that of cluster $\{A, D, G, H\}$. Of course instead of $\hat{\eta}_B$ we might have written $\hat{\eta}_E$ or $\hat{\eta}_I$ (since $\hat{\eta}_B = \hat{\eta}_E = \hat{\eta}_I$) because schools B, E and I belong to the same cluster. Similarly the odds of a Low instead of High grade in clusters $\{F, J\}$, $\{K\}$ and $\{L\}$ are respectively 1.8610, 1.4417, 2.3485 times that of cluster $\{A, D, G, H\}$. These conclusions are in perfect agreement with those presented by Leonard and Novick (1986, p 47).

5 Discussion

In this paper we have presented a new method for the Bayesian analysis of two-way contingency tables based on a row effects model. This model is especially appropriate when one of the variables, say the column, is measured on an ordinal scale. The main idea is to assume a multiplicative structure for the interaction term in the log-linear expansion consisting of a row effect parameter and a column score, the latter being fixed and monotone increasing with respect to the arrangement of the column levels. In this way row effects acquire a simple and intuitive interpretation: differences of row effects represent, on a log scale, odds ratios relative to any adjacent pair of levels for the column variable.

We focused on methods for clustering the row effects. Our interpretation of clustering is particularly stringent in this context: namely two row effects are declared to belong to the same cluster when they are (stochastically) equivalent. In this way we simplify the interaction structure of the model and enhance its interpretation in terms of odds that are constant within each cluster. When applied to synthetic, as well as real data sets, our method shows attractive features and a very good performance.

Another useful interpretation of the equality between two row effects relates to the notion of merging in contingency tables, see Wermuth and Cox (1998) and Dellaportas and Tarantola (2005). Specifically, if $\eta_k = \eta_i$ then all log odds for adjacent categories of variable B in the $2 \times b$ subtable relative to rows $\{i, k\}$ are equal to zero. This entails independence properties that allow for a merging of the above rows. The issues of merging in the context of RC-models is dealt with in Kateri and Iliopoulos (2003).

Our approach is carried-out in a Bayesian framework and is entirely model-based. The latter feature marks the difference with previous work in the area that was mostly *ad hoc* and descriptive, as described in the discussion of the Marine corps data set in Section 4.3. In particular, the clustering component of our method is based on a product partition model (PPM) that allows to deal simultaneously with models of varying dimensions (corresponding to alternative

clustering structures), without resorting to elaborate MCMC techniques, such as Reversible Jump. By a suitable choice of the cohesion function in our PPM, we are able to draw on previous research in the area of computational Bayesian nonparametrics using mixture of Dirichlet process priors, in order to construct an MCMC sampler that simultaneously explores the space of parameters and models. The output of our sampler can be used for a variety of purposes: in particular estimating features of the posterior distribution of the parameters of interest as well as calculating posterior probabilities on the space of partitions. The latter allows to implement a Bayesian Model Averaging analysis which consists in a weighted combination of conditional inferences (on each partition) with weights equal to the posterior probabilities of the corresponding partitions. For a Bayesian analysis of row and column effects models (RC models) see Evans *et al.* (1993) and Kateri *et al.* (2005).

Possible future directions of research, along the lines indicated in this paper, include the specification of alternative cohesion functions for the product distribution, as well as the corresponding MCMC sampler. A more challenging, and potentially very useful, endeavour would be to extend the scope of our methodology to contingency tables wherein also the row variable is ordinal; this would entail dealing with contiguous partitions, see Section 4.2 for some preliminary remarks. Finally, our clustering method can be naturally extended to row and column (RC) effects models; for a Bayesian analysis of RC models, see Evans *at al.* (1993) and Kateri *et al.* (2005).

Ackwnowledgements

The research of the first two authors was partially supported by MIUR, Rome (PRIN 2003138887 and PRIN 2005132307), and the University of Pavia. The authors are thankful to Stefano Sampietro and Luigi Spezia for helpful discussions.

References

Antoniak, C.E. (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.*, **2**, 1152-1174.

Agresti, A. and Chuang, C. (1989) Model-based Bayesian methods for estimating cell proportions in cross-classification tables having ordered categories. *Comput. Statist. Data Anal.*, **7**, 245-258.

Agresti, A. (1999) Modelling ordered categorical data: recent advances and future challenges. *Statist. Med.*, **18**, 2191-2207.

Agresti, A. (2002) *Categorical Data Analysis*. New York: Wiley.

Barry, D. and Hartigan, J.A. (1992) Product partition models for change point problems. *Ann. Statist.*, **20**, 260-279.

Brink, A. and Smith, A.F.M. (1996) Bayesian modelling of the association in contingency tables. In *Statistical Modelling: Proceedings of the 11th International Workshop On Statistical Modelling*, Forcina, A., Marchetti, G. M., Hatzinger, R. and Galmacci, G. (eds.), Orvieto, 95-103.

Brooks, S.P. and Gelman, A. (1998) General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.*, **7**, 434-455.

Bush, C.A. and MacEachern, S.N.(1996) A semiparametric Bayesian model for randomised block designs. *Biometrika*, **83**, 275-285

Clogg, C. C. and Shihadeh, E.S. (1994) *Statistical Methods for Ordinal Variables*. Thousand Oaks: Sage.

Congdon, P. (2001) *Bayesian Statistical Modelling*. Chichester: Wiley.

Consonni, G. and Veronese, P. (1995) A Bayesian method for combining results from several binomial experiments. *J. Am. Statist. Assoc.*, **90**, 935-944.

Crowley, E. M. (1997) Product partition models for normal means. *J. Am. Statist. Assoc.*, **92**, 192-198.

Dellaportas, P. and Tarantola, C. (2005) Model determination for categorical data with factor level merging. *J. R. Statist. Soc B*, **67**, 269-283.

Denison, D.G.T., Adams, N.M., Holmes, C.C. and Hand, D.J. (2002) Bayesian partition modelling. *Comput. Statist. Data Anal.*, **38**, 475-485.

Escobar, M.D. and West, M. (1995) Bayesian density estimation and inference using mixtures. *J. Am. Statist. Assoc.*, 89, 268-277.

Etzioni, R.D., Fienberg, S.E., Gilula, Z. and Haberman, S.J. (1994) Statistical models for the analysis of ordered categorical data in public health and medical research. *Statist. Meth. Medic. Res.*, **3**, 179-204.

Evans, M., Gilula, Z., Guttman, I. and Swar, T. (1993) Computational issues in the Bayesian analysis of categorical data: log-linear and Goodman's RC model. *Statist. Sinica*, **3**, 391-406.

Evans, M., Gilula, Z., Guttman, I. and Swar, T. (1997) Bayesian analysis of stochastically ordered distributions of categorical variables. *J. Am. Statist. Assoc.*, **92**, 208-214.

Geyer C. J. (1992) Practical Markov chain Monte Carlo (with Discussion). *Statist. Sci.*, **7**, 473-511.

Goodman, L. A. (1979) Simple models for the analysis of association in cross-classifications having ordered categories. *J. Am. Statist. Assoc.*, **74**, 537-552.

Goodman, L. A. (1983) The analysis of dependence in cross-classification having ordered categories, using log-linear models for frequencies and log-linear models for odds. *Biometrics* , **39**, 149-160.

Hartigan, J.A. (1990) Partition models. *Commun. Statist. Theory Meth.*, **19**, 2745-2756.

Jain, S. and Neal, R. M. (2004) A split-merge Markov Chain Monte Carlo procedure for the Dirichlet process mixture model, *J. Comput. Graph. Statist.*, **13**, 158-182

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

Kateri, M. and Iliopoulos, G. (2003) On collapsing categories in two-way contingency tables, *Statistics*, **37**, 443-455.

Kateri, M., Nicolaou, A. and Ntzoufras, I.(2005) Bayesian inference for the RC(m) association model, *J. Comput. Graph. Statist.*, **14**, 116-138

Lang, J.B. (1999) Bayesian ordinal and binary regression models with a parametric family of mixture links. *Computat. Statist. Data Anal.*, **31**, 59-87.

Leonard, T. and Novick, M. R. (1986) Bayesian full rank marginalization for two-way contingency tables. *J. Educ. Statist.*, **11**, 33-56.

Loschi, R. H., Cruz, F.R.B. , Iglesias, P.L. and Arellano-Valle, R.B. (2003) A Gibbs sampling scheme to the product partition model: an application to change-point problems. *Computers Operat. Res.*, **30** , 463 - 482.

MacEachern, S.N. and Müller, P. (1998) Estimating mixture of dirichlet process models. *J. Computat. Graph. Statist.*, **7**, 223-238.

MacEachern, S.N. and Müller, P. (2000) Efficient MCMC schemes for robust model extensions using encompassing Dirichlet process mixture models. In *Robust Bayesian Analysis*, Rios Insua, D. and Ruggeri, F. (eds), New York: Springer, 295-315.

Malec, D. and Sedransk, J. (1992) Bayesian methodology for combining the results from different experiments when the specifications for pooling are uncertain. *Biometrika*, **79**, 593-601.

Powers, D.A. Xie, Y. (2000). *Statistical Methods for Categorical Data Analysis*, San Diego: Academic Press.

Quintana, F.A. and Iglesias, P.L. (2003) Bayesian clustering and product partition models. *J. Roy. Statist. Soc. B*, **65**, 557-574.

Smith, B. (2001) Bayesian Output Analysis Program: version 1.0.0. Dept. of Biostatistics. University of Iowa, USA.
(available at *http://www.public-health.uiowa.edu/boa*)

Wermuth, N. and Cox, D.R. (1998) On the application of conditional independence to ordinal data. *Inter. Statist. Review*, **66**, 181-199.

West, M. Müller, P. and Escobar, M.D. (1994) Hierarchical priors and mixture models, with application: in regression and density estimation, In *Aspects of Uncertainty: a Tribute to D.V. Lindely*, Freeman, P. R. and Smith, A.F.M. (eds.), Chichester: Wiley, 263-386.

Table 1: Simulated data

A	B				
	b_1	b_2	b_3	b_4	b_5
a_1	12	90	33	90	148
a_2	245	403	33	20	7
a_3	20	148	55	148	245
a_4	55	90	7	4	2
a_5	20	148	55	148	245
a_6	148	245	20	12	4
a_7	20	148	55	148	245
a_8	55	90	7	4	2
a_9	33	245	90	245	403
a_{10}	90	245	33	33	20

Table 2: Posterior model probabilities for the simulated data

Partition	Posterior probabilities
{1,3,5,7,9} {2,4,6,8}{10}	0.9378 (16)[a]
{1,3,5,7,9} {2,6,8} {4}{10}	0.0111 (06)
{1,3,5,7,9} {2,4,6} {8}{10}	0.0111 (06)
{1,3,5,7,9} {2,4,8} {6}{10}	0.0100 (39)

[a]Figures in brackets are Monte Carlo standard errors $\times 10^4$.

Table 3: Premarital sex data

	Teenage birth control			
Premarital sex	Strongly disagree	Disagree	Agree	Strongly agree
	81	68	60	38
Always wrong	(42.41)[a]	(51.21)	(86.42)	(66.95)
	[74.93][b]	[66.78]	[73.10]	[32.57]
	24	26	29	14
Almost always wrong	(15.96)	(19.28)	(32.54)	(25.20)
	[27.88]	[25.11]	[27.79]	[12.53]
	18	41	74	42
Wrong only sometimes	(30.05)	(36.28)	(61.23)	(47.43)
	[24.13]	[35.99]	[65.93]	[49.17]
	36	57	161	157
Not wrong at all	(70.57)	(85.22)	(143.80)	(111.40)
	[32.32]	[64.40]	[157.60]	[157.00]

[a]Fitted counts under independence model

[b]Expected posterior counts under the highest-posterior-probability PPM

Table 4: Posterior model probabilities for the Premarital sex data

Partition	Posterior probabilities
{1,2} {3}{4}	0.9054 (65)[a]
{1} {2} {3}{4}	0.0946 (65)

[a]Figures in brackets are Monte Carlo standard errors $\times 10^4$.

Table 5: Premarital sex data, posterior expectations and standard deviations of parameter estimates under the highest-posterior-probability PPM

Parameter		Posterior Expectation	Posterior SD
Premarital sex	Always wrong	1.0651	0.0427
Premarital sex	Almost always wrong	0.0883	0.0571
Premarital sex	Wrong only sometimes	-0.5782	0.0482
Premarital sex	Not wrong at all	-0.5752	0.0379
Teenage Birth Control	Strongly disagree	2.0435	0.0479
Teenage Birth Control	Disagree	0.7280	0.0440
Teenage Birth Control	Agree	-0.3814	0.0370
Teenage Birth Control	Strongly agree	-2.3901	0.0434
Row effect (η_1)	Always wrong	1.2031	0.0029
Row effect (η_2)	Almost always wrong	1.2031	0.0029
Row effect (η_3)	Wrong only sometimes	1.7156	0.0040
Row effect (η_4)	Not wrong at all	2.0051	0.0028

Table 6: Marine corps data, collapsed table

School	Grade	
	High	Low
A	475 (456.70)[a] [472.44] [b]	480 (498.29) [482.90]
B	202 (155.90) [199.63]	124 (170.10) [126.60]
C	708 (615.47) [707.78]	579 (671.53) [579.45]
D	90 (81.30) [87.05]	80 (88.70) [83.12]
E	89 (70.78) [90.13]	59 (77.22) [58.07]
F	229 (313.71) [227.23]	427 (342.29) [429.17]
G	410 (403.14) [417.00]	433 (439.86) [426.31]
H	95 (91.82) [95.31]	97 (100.18) [96.86]
I	109 (86.8) [109.72]	71 (93.92) [70.55]
J	78 (11.43) [80.45]	155 (121.57) [152.78]
K	81 (95.64) [80.50]	119 (104.35) [119.75]
L	135 (219.03) [135.24]	323 (238.97) [323.94]

[a]Fitted counts under independence model

[b]Expected posterior counts under the highest-posterior-probability PPM

Table 7: Posterior model probabilities for the Marine corps data

Partition	Posterior probabilities
{ADGH} {BEI}{FJ} {C}{K}{L}	0.5775 (160)[a]
{AGH} {BEI} {FJ}{CD}{KL}	0.1979 (120)
{ADGH} {BEI} {FJK}{C}{L}	0.0414 (79)
{ADGH} {BEI} {JL}{C}{F}{K}	0.0236 (58)
{ADGH} {BI} {EC}{FJ}{K}{L}	0.0177 (37)
{AGH} {CDEI} {FL}{B}{F}{K}	0.0176 (60)
{AGH} {BEI} {CD}{FJK} {L}	0.0153 (39)
{AG} {BEI} {CDH}{FJ} {K}{L}	0.0150 (30)
{AGH} {BE} {CDI}{FJ} {K}{L}	0.0121 (34)
{AGH} {BI} {CDE}{FJ} {K}{L}	0.0114 (29)

[a]Figures in brackets are Monte Carlo standard errors $\times 10^4$.

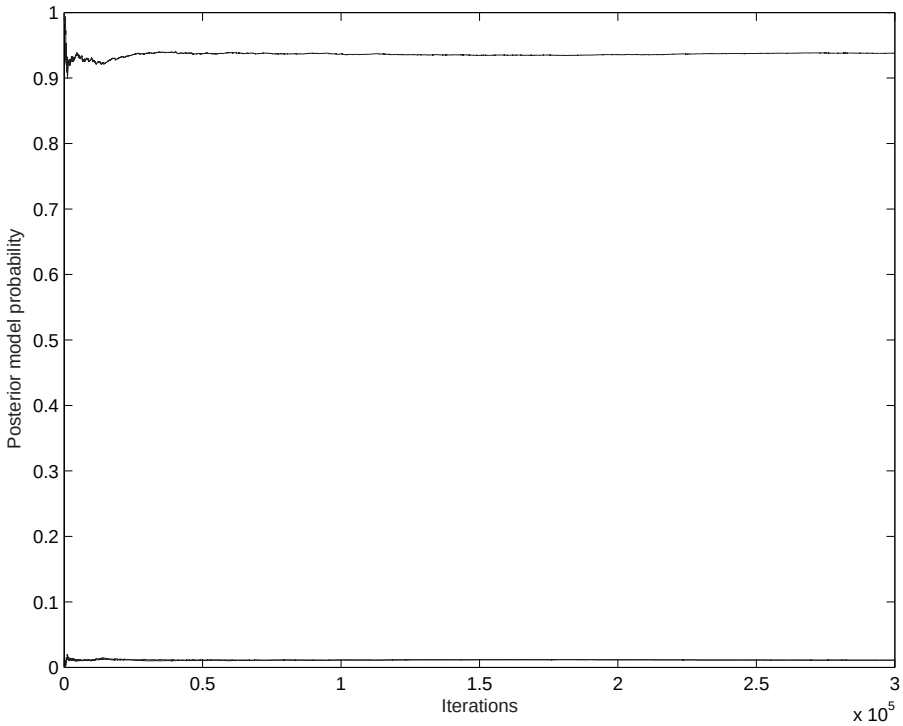


Figure 1: Simulated data, ergodic means for the posterior model probabilities

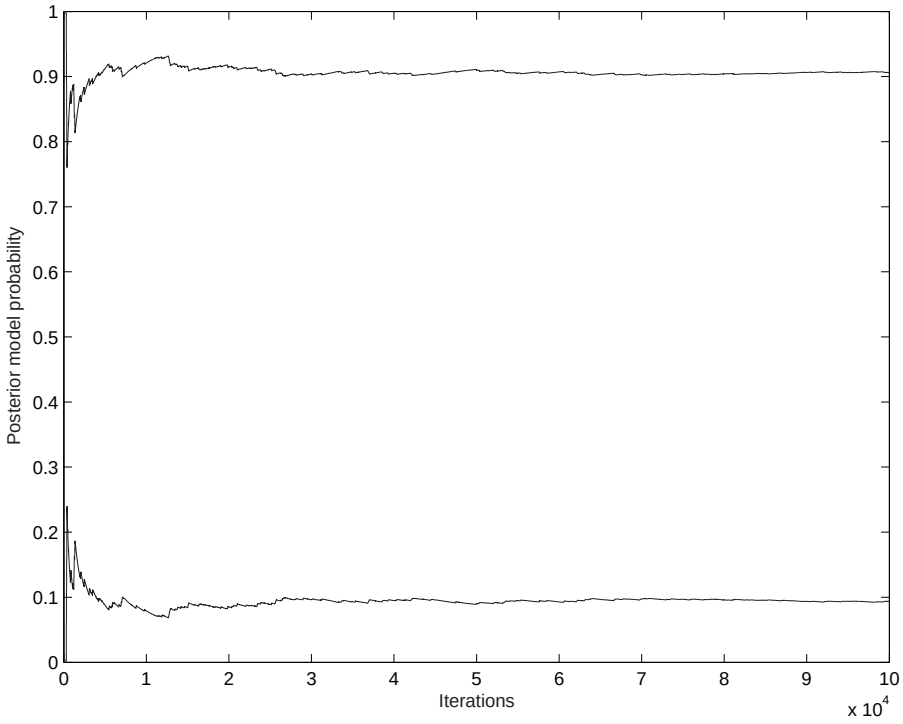


Figure 2: Premarital set data, ergodic means for the posterior model probabilities for the two most probable models

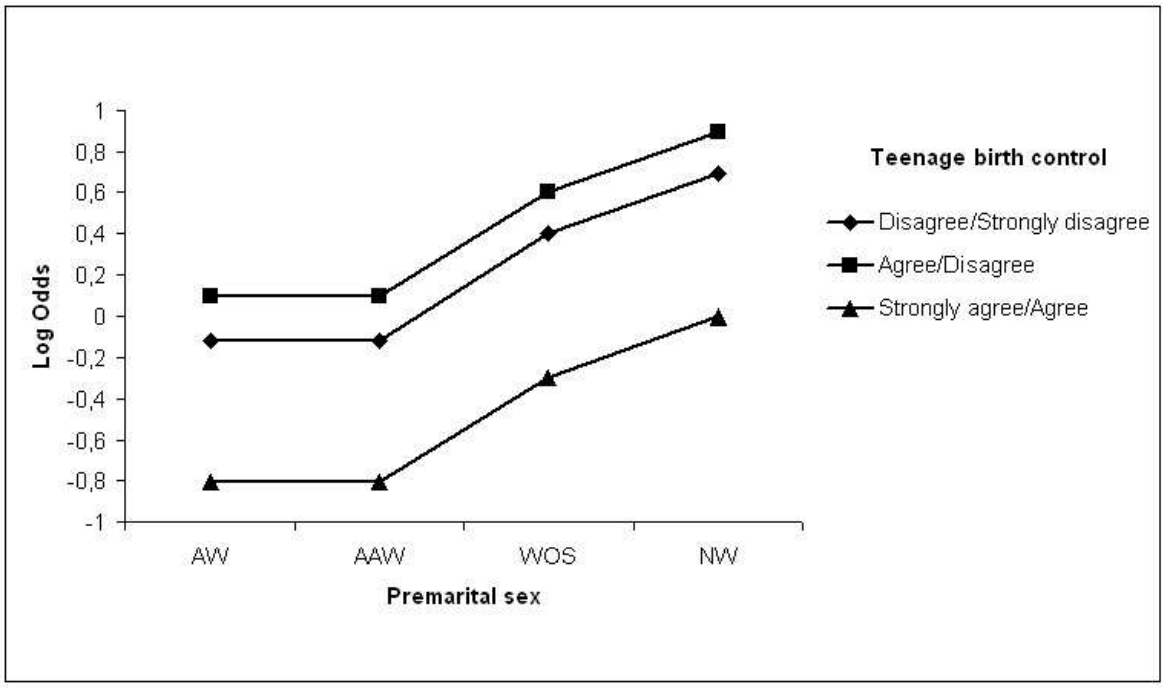


Figure 3: Premarital set data, predicted logits for adjacent column categories