

A Bayesian Test of Independence for Two-way Contingency Tables Under Cluster Sampling

by

Dilli Raj Bhatta

A Thesis

Submitted to the Faculty

of

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Doctor of Philosophy

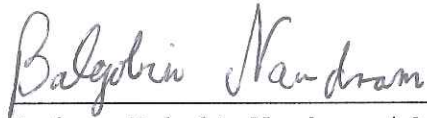
in

Mathematical Sciences

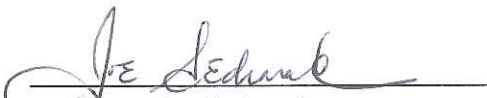


December 04, 2012

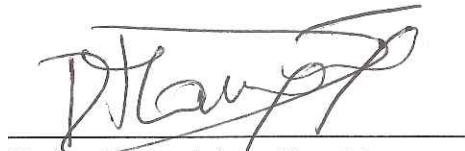
APPROVED:



Professor Balgobin Nandram, Advisor
Department of Mathematical Sciences
Worcester Polytechnic Institute



Professor Joe Sedransk
Department of Statistics
Case Western Reserve University



Professor Dominique Haughton
Department of Mathematical Sciences
Bentley University



Dr. Jai Won Choi
Statistical Consultant, Meho Inc.
9504 Mary Knoll Dr., Rockville MD 20850



Professor Hasanjan Sayit
Department of Mathematical Sciences
Worcester Polytechnic Institute

Contents

1	Introduction	1
1.1	Literature Review	3
1.1.1	Rao Scott Chi-Square Test	3
1.1.2	Brier (1980) Model	7
1.1.3	Geenens and Simar (2010, 2011)	8
1.2	Surrogate Sampler, Nandram (2007)	9
1.3	Bayes Factor for a Test of Independence	10
1.4	Applications	11
1.4.1	Third International Mathematics and Science Study (TIMSS 1995)	11
1.4.2	Trend in International Mathematics and Science Study (TIMSS 2007)	12
1.5	Plan of Dissertation	13
2	A Test of Independence Without Covariates	14
2.1	Hierarchical Bayesian Model	15
2.2	Computations, Bayes Factor and Specifications	18
2.2.1	Computations	18
2.2.2	Bayes Factor	20
2.2.3	Specifications	22
2.3	Numerical Analysis	23
2.3.1	Illustrative Example	23
2.3.2	Simulation Study	26
2.4	Concluding Remarks	29
3	A Test of Independence With Covariates	35
3.1	A Random Effect Multinomial Logistic Regression Model	36
3.1.1	The Joint Posterior Density	37
3.1.2	Computation	43
3.1.3	Assessing the Model Fit	46
3.1.4	Surrogate Cluster Sample Without Covariates	47
3.2	Cluster Model Without Covariates	48
3.2.1	Hierarchical Bayesian Model	48
3.2.2	Computation	49
3.3	Bayes Factor	51
3.4	Applications	52
3.5	Simulation Study	55

3.6	Power Function	57
3.7	Concluding Remarks	58
4	Concluding Remarks	69
4.1	Contribution in Methodology	69
4.2	Interpretations of Surrogate Sampling Table and discussion	71
4.3	Future Work	75
4.3.1	Stratified Two-stage Cluster Sampling	75
4.3.2	Introduction of Survey Weights	75
4.3.3	Sampling Zero Problem	77
4.4	My Accomplishments	79
4.4.1	Mortality Curve Fitting	79
4.4.2	Selection Bias	80
4.4.3	Sparse Two-Way Contingency Tables	80
A	Joint Posterior Density	81
B	A Property of the Gamma Distribution	83
C	Mode of a Kernel Density Estimator	84
D	Joint Posterior Density: A Simplification	85
E	Proof that $\int_{\infty}^{\infty} h(\nu_i)d\nu_i$ is finite.	87
F	Cluster Tables for Examples E1-E6 using TIMSS 2007 Data	89

List of Figures

2.1	Plots of the empirical densities of the log-Bayes factors for the eight strata in the third grade example	33
2.2	Simulation: Plots of the empirical densities of the log-Bayes factors at twelve design points. The symbols are correlation (solid: $\rho = .01$, dotted: $\rho = .10$, long dashed: $\rho = .30$), association (independence: ind=1 and dependence: ind=2) and table density (Δ)	34
3.1	Plot of the empirical densities of the log-Bayes factors for the simulation with covariates when ind=1.0, ind=1.2 and ind=1.4. In the legend on the top right side, the first and second values of the pair represent the intracluster correlation (ρ) and the cluster size (ℓ) respectively.	67
3.2	Plot of the estimated power function of the test	68

List of Tables

2.1	Features of the total table for each of the eight examples	30
2.2	Bayesian Design effects for each cell by example	30
2.3	Comparison of the log-Bayes factor with the p-values by example	31
2.4	Sensitivity analysis of the log-Bayes factor with respect to $\tau_s, s = 1, \dots, 9$, by region (reg) and example	31
2.5	Simulation: Comparison of p-values and log-Bayes factor	32
3.1	Features of the total table for each of the six examples	59
3.2	Posterior estimate of the parameters under multinomial logistic regression	60
3.3	Summary of the log-Bayes factor	62
3.4	Bayesian design effects for each cell by example	62
3.5	Study of the effects of covariates on the test of independence	63
3.6	Study the effects of covariates under simple random sampling	64
3.7	Sensitivity analysis of the log-Bayes factor with respect to a and b in the prior of σ^2 , by examples	65
3.8	Simulation: summary of the log-Bayes factor for the cluster model with covariates (MWC) and without covariates (MWOC)	66
3.9	Simulation: summary of the log-Bayes factor for the cluster model with covariates (MWC) and without covariates (MWOC)	66
4.1	Comparison of the observed total table and the surrogate total table by example	73
4.2	Comparison of inference from the observed total table and the surrogate total table by example	74

Acknowledgments

I would like to extend my sincere gratitude to my advisor, Balgobin Nandram. It has been an honor to be his student. I appreciate all his contributions in providing the ideas and time throughout my Ph.D. career. Without his support, guidance, care and promptness in research and writing the dissertation, I would not be able to complete my work on time. I am also very grateful to all of my committee members for their time and guidance. Their brilliant comments and suggestions have helped to make this dissertation more decent.

I would like to thank the department for providing the financial support for my entire period at WPI, all its faculty and staff for their help and kindness. I would like to thank all my friends in the department. I would also like to thank Professor Dhiman Bhadra for providing a six months' research assistantship, and the National Institute of Statistical Sciences (NISS) for providing a three months' summer internship which provided an opportunity to gain some experiences while working in real problem.

Finally, I would like to thank my wife for her understanding, support and care; my two lovely daughters whose arrival in this world brought light into my life; my cousin, Ghanshyam Bhatt, whose encouragement, support and guidance brought me to this stage. Special thanks also go to all of the members of the Bhatta families, my in-laws and all other relatives whose emotion, belief and support provided the strength within me to achieve this goal.

*This dissertation is dedicated to
my grandparents in heaven.*

Abstract

We consider a Bayesian approach to the study of independence in a two-way contingency table obtained from a two-stage cluster sampling design. We study the association between two categorical variables when (a) there are no covariates and (b) there are covariates at both unit and cluster levels. Our main idea for the Bayesian test of independence is to convert the cluster sample into an equivalent simple random sample which provides a surrogate of the original sample. Then, this surrogate sample is used to compute the Bayes factor to make an inference about independence.

For the test of independence without covariates, the Rao-Scott corrections to the standard chi-squared (or likelihood ratio) statistic were developed. They are “large sample” methods and provide appropriate inference when there are large cell counts. However, they are less successful when there are small cell counts. We have developed the methodology to overcome the limitations of Rao-Scott correction. We have used a hierarchical Bayesian model to convert the observed cluster samples to simple random samples. This provides the surrogate samples which can be used to derive the distribution of the Bayes factor to make an inference about independence. We have used a sampling-based method to fit the model.

For the test of independence with covariates, we first convert the cluster sample with covariates to a cluster sample without covariates. We use multinomial logistic regression model with random effects to accommodate the cluster effects. Our idea is to fit the cluster samples to the random effect models and predict the new samples by adjusting with the covariates. This provides the cluster sample without covariates. We then use a hierarchical Bayesian model to convert this cluster sample to a simple random sample which allows us to calculate the Bayes factor to make an inference about independence. We use Markov chain Monte Carlo methods to fit our models.

We apply our first method to the Third International Mathematics and Science Study (1995) for third grade U.S. students in which we study the association between the mathematics test scores and the communities the students come from, and science test scores and the communities the students come from. We also provide a simulation study which establishes our methodology as a viable alternative to the Rao-Scott approximations for relatively small two-stage cluster samples.

We apply our second method to the data from the Trend in International Mathematics and Science Study (2007) for fourth grade U.S. students to assess the association between the mathematics and science scores represented as categorical variables and also provide the simulation study. The result shows that if there is strong association between two categorical variables, there is no difference between the significance of the test in using the model (a) with covariates and (b) without covariates. However, in simulation studies, there is a noticeable difference in the significance of the test between the two models when there are borderline cases (i.e., situations where there is marginal significance).

Chapter 1

Introduction

Analysis of categorical data presented in a two-way contingency table is a well known problem. In order to analyze such tables obtained from simple random sampling, the Pearson chi-squared and the likelihood ratio tests are commonly used for testing association between two categorical variables. These tests depend on the assumption that the data in the table follow the multinomial distribution. This dissertation is focused on analyzing such tables when the data are obtained from a two-stage cluster sampling design with simple random sampling at both stages. We note that Nandram and Sedransk (1993) provided a Bayesian procedure to obtain inference about a finite population proportion under two-stage cluster sampling, see also Nandram (1998). We study the association between two categorical variables when (a) there are no covariates and (b) there are covariates at both unit and cluster levels. We use a Bayesian test of independence to study the association between two categorical variables. Our main idea for the Bayesian test of independence is to convert the cluster sample into an equivalent simple random sample which provides a surrogate of the original sample. Then, we use the Bayes factor to make an inference about independence.

It is pertinent to give some background information. Let $\{n_{jk}, j = 1, \dots, r, k = 1, \dots, c\}$ denote the cell counts in an $r \times c$ contingency table and let $n = \sum_{j=1}^r \sum_{k=1}^c n_{jk}$ denote the total sample size. The marginal totals for the j^{th} row and k^{th} column are respectively $n_{j\cdot} = \sum_{k=1}^c n_{jk}, j = 1, \dots, r$, and $n_{\cdot k} = \sum_{j=1}^r n_{jk}, k = 1, \dots, c$. Let $S = rc$ denote the total number of cells and π_{jk} the cell probability for the $(j, k)^{\text{th}}$ cell, where we assume that $\pi_{jk} > 0$ for all j and k and $\sum_{j=1}^r \sum_{k=1}^c \pi_{jk} = 1, p_j = \sum_{k=1}^c \pi_{jk}$ and $q_k = \sum_{j=1}^r \pi_{jk}$. The independence hypothesis states that $\pi_{jk} = p_j q_k, j = 1, \dots, r, k = 1, \dots, c$, where $\sum_{j=1}^r p_j = \sum_{k=1}^c q_k = 1$.

Let $\tilde{\pi}_{jk} = n_{j \cdot n \cdot k} / n^2$ denote the maximum likelihood estimates of the π_{jk} under the hypothesis of independence of the two categorical variables. Letting $\hat{\pi}_{jk} = n_{jk} / n$, the Pearson chi-squared and the likelihood ratio statistics are, respectively,

$$X^2 = n \sum_{jk} (\hat{\pi}_{jk} - \tilde{\pi}_{jk})^2 / \tilde{\pi}_{jk}, \quad G^2 = 2n \sum_{jk} \hat{\pi}_{jk} \log(\hat{\pi}_{jk} / \tilde{\pi}_{jk}),$$

where $\tilde{\pi}_{jk}, j = 1, \dots, r, k = 1, \dots, c$, are assumed to be positive and arising from simple random sampling. It is well known that under the null hypothesis of independence, both X^2 and G^2 have equivalent asymptotic ($n \rightarrow \infty$ with S fixed) chi-squared distributions with $(r - 1)(c - 1)$ degrees of freedom.

The chi-squared distribution of the Pearson chi-squared and likelihood ratio test statistic results from simple random sampling assumption. However, these tests are not appropriate with complex survey designs; in fact, this results in incorrect p-values. For example, when there is a clustering effect, the units in a cluster are, in general, positively correlated. Due to intracluster correlation, the usual multinomial sampling scheme is no longer appropriate because of the violation of the assumptions in the multinomial distribution. Specifically, the standard chi-squared or likelihood ratio test can fail. If a procedure based on simple random sampling rather than cluster sampling is used to test for independence, the p-value can be too small (or X^2, G^2 values too big), resulting in significant evidence against the null hypothesis when there may be no such evidence. For a complex sample design (e.g., two-stage cluster sampling, stratified multistage cluster sampling, etc.), both X^2 and G^2 have ‘skewed’ distributions.

Rao and Scott (1981, 1984) obtained the design-adjusted version of the Pearson chi-squared test using the simple correction to the standard X^2 and G^2 statistics for the test of independence in a two-way contingency table arising from any complex sampling design. The corrections are based on normal approximations and moment-matching principles and they are obtained through *design effects*. A design effect is the ratio of the variance of a statistic under a complex sampling design to the variance of the statistic under simple random sample. For two-stage cluster sampling, these design effects can be much larger than one, under the assumption of positive correlation within the cluster thereby having a large impact on the

standard chi-squared statistic. The tests perform well for large complex surveys, but for smaller surveys, Rao-Scott corrections are not accurate partly, because the chi-squared test is inaccurate.

Brier (1980) used a Multinomial-Dirichlet distribution to model the distribution of counts in a two-way contingency table under cluster sampling. He has shown how to make an adjustment to the usual goodness-of-fit statistic by using Multinomial-Dirichlet model. However, the model provides the same design effects for all estimators of cell probabilities of a two-way table.

When there is a set of the characteristics (e.g., covariates) at unit level and/or cluster level, the problem becomes more practical. This is because these covariates are likely to be associated with the two cross-classified categorical variables and can influence their association. If we simply ignore the effect of covariates, the test can be misleading. Under simple random sampling design, Geenens and Simar (2010, 2011) developed nonparametric and semiparametric methods for conditional independence in two-way contingency tables.

In Chapter 1, we give some detailed background information. In Section 1.1, we review some closely related literature. In Section 1.2 we briefly discuss the concept of surrogate sampling. In Section 1.3 we discuss the Bayes factor. Finally, in Section 1.4 we discuss two applications which use our methods.

1.1 Literature Review

In this section, we review some of the existing literature on the test of independence in a two-way contingency table for complex sampling designs and the conditional test of independence between two categorical variables given a vector X of continuous covariates under simple random sampling.

1.1.1 Rao Scott Chi-Square Test

Rao and Scott (1981, 1984) show that, under very general complex designs, X^2 and G^2 are still asymptotically equivalent. Let \hat{P} denote a consistent estimator of the cell probabilities and $V = \text{cov}(\hat{P})$, where \hat{P} can be very complex as it can involve survey weights and other

design features. Assuming that the central limit theorem holds, Rao and Scott (1981, 1984) show that $X^2 = \sum_{i=1}^{\kappa} \delta_i Z_i^2 = G^2$, where the $Z_i, i = 1, \dots, \kappa$ are independent standard normal random variables and $\delta_i, i = 1, \dots, \kappa$, are the positive eigenvalues associated with V and the design matrix. The δ_s are known as *generalized design effects*, a phrase originally coined by Rao and Scott (1981). Let \hat{V} be an estimator of V . If the entire data set is available, \hat{V} can be obtained using linearization or a resampling method (e.g., bootstrap or jackknife). Let $\hat{\delta}_i$ be the consistent estimators of $\delta_i, i = 1, \dots, \kappa$, and $\hat{\bar{\delta}}$ be the same for $\bar{\delta} = \sum_{i=1}^{\kappa} \delta_i / \kappa$. Then, the effective sample size (Fellegi, 1980) in the complex survey equivalent to the simple random sample is $\tilde{n} = n / \bar{\delta}$, and the Rao and Scott (1981) adjusted X^2 and G^2 are

$$\bar{X}^2 = \tilde{n} \sum_{jk} (\hat{\pi}_{jk} - \tilde{\pi}_{jk})^2 / \tilde{\pi}_{jk}, \quad \bar{G}^2 = 2\tilde{n} \sum_{jk} \hat{\pi}_{jk} \log(\hat{\pi}_{jk} / \tilde{\pi}_{jk}).$$

For a two-stage cluster sampling design, \tilde{n} can be much smaller than n depending on the intra-cluster correlation. Rao and Scott (1981) obtained a first order approximation by matching first moments of the distributions, and a second order approximation by matching the first two moments of the distributions using Satterthwaite's procedure, ignoring the sampling variation in \hat{V} .

A third approximation, an adjustment which uses the degree of freedom in the variance estimate to account for sampling variation in \hat{V} and other parameters, is more accurate than the first two methods; see Thomas and Rao (1987), Rao and Thomas (1989) and Thomas, Singh and Roberts (1996). However, the first order approximation is typically used in practice (e.g., SAS Proc Surveyfreq Version 9.2) and can be calculated using information on the standard errors of the cell probabilities and marginal proportions which are generally available (e.g., see Bedrick 1983). Thus Rao-Scott corrections are very useful and practical for large complex surveys.

However, for smaller complex surveys (i.e., when expected cell counts are less than 5), the asymptotic distributions of both \bar{X}^2 and \bar{G}^2 can be grossly incorrect and hence their applicability is questionable. In this case, the Rao-Scott corrections are not appropriate because they are not constructed to deal with small expected cell counts.

We discuss below in a very simple way how the first and second order corrections are

obtained.

First Order Correction

In an $r \times c$ contingency table obtained from a complex survey, the test statistics X^2 and G^2 do not follow a χ_b^2 , $b = (r - 1)(c - 1)$ distribution under the null hypothesis of independence. But both statistics have skewed distributions, and a multiple of X^2 or G^2 may approximately follow a χ_b^2 distribution. That is, approximately

$$cX^2 \sim \chi_b^2,$$

where c is to be determined. Under the first order correction, the mean of the test statistic is matched with the mean of a χ_b^2 random variable. This gives $c = b/E(X^2)$ and thus $X^2/\{E(X^2)/b\} \sim \chi_b^2$. Because $X^2 \approx \sum_{i=1}^b \delta_i Z_i^2$, where Z_i are independent standard normal random variables, we get $E(X^2) = \sum_{i=1}^b \delta_i$. Thus,

$$E(X^2)/b = \sum_{i=1}^b \delta_i/b = \bar{\delta},$$

where $\bar{\delta}$ is called a design correction. Therefore, $X^2/\bar{\delta} \sim \chi_b^2$ is a first order corrected test statistic.

Second Order Correction

Under the second order correction, the two moments (mean and variance) of the test statistic are matched with the mean and the variance of $\chi_{k_1^*}^2$, k_1^* is unknown (as done by Satterthwaite, 1946). Because $cX^2 \sim \chi_{k_1^*}^2$, matching the moments we get $c = 2E(X^2)/V(X^2)$ and $k_1^* = 2\{E(X^2)\}^2/V(X^2)$, and because $X^2 \sim \sum_{i=1}^b \delta_i Z_i^2$, we get $E(X^2) = \sum_{i=1}^b \delta_i$ and $V(X^2) = \sum_{i=1}^b \delta_i^2$.

Third Order Correction

Under the third order correction $cX^2 \sim F_{k_1^*, k_2^*}$, where $k_2^* = k_1^* \nu$ with $\nu = \text{rank}(\hat{V})$. Note that typically $\nu = \#$ of primary sampling units (psu's) - $\#$ of strata which may be relatively small even in big surveys (see Scott, 2007). However, if it is only cluster sampling design, ν would be equal to b .

Now, we briefly describe the Rao and Scott (1981) approach for the test of independence

in a two-way table. Suppose there are r rows and c columns, the hypothesis of interest is

$$H_0 : h_{ij}(p) = p_{ij} - p_{i+}p_{+j}, \quad i = 1, \dots, r-1; \quad j = 1, \dots, c-1,$$

where $p = (p_{11}, p_{12}, \dots, p_{rc-1})'$, $p_{i+} = \sum_{j=1}^c p_{ij}$ and $p_{+j} = \sum_{i=1}^r p_{ij}$, and p_{ij} is the population proportion in the $(i, j)^{\text{th}}$ cell. Let n_{ij} be the observed cell count in the $(i, j)^{\text{th}}$ cell and $\hat{n}_{ij} = n\hat{p}_{i+}\hat{p}_{+j}$ be the corresponding expected count under independence. Letting $\hat{p}_{ij} = n_{ij}/n$, where $n = \sum_{i=1}^r \sum_{j=1}^c n_{ij}$, the sample size, the usual Pearson statistic for testing H_0 is

$$\begin{aligned} X^2 &= \sum_{i=1}^r \sum_{j=1}^c (n_{ij} - \hat{n}_{ij})^2 / \hat{n}_{ij} \\ &= \sum_{i=1}^r \sum_{j=1}^c (n_{ij} - n\hat{p}_{i+}\hat{p}_{+j})^2 / n\hat{p}_{i+}\hat{p}_{+j} \\ &= n \sum_{i=1}^r \sum_{j=1}^c (\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j})^2 / \hat{p}_{i+}\hat{p}_{+j}, \end{aligned}$$

which can be rewritten as

$$X^2 = n\tilde{h}(\hat{p})'(\hat{P}_r^{-1} \otimes \hat{P}_c^{-1})\tilde{h}(\hat{p}). \quad (1.1)$$

Here, \otimes denotes the direct matrix product, \hat{p}_{ij} is the estimate of p_{ij} under the sampling design $p(s)$, $\tilde{h}(\hat{p})$ is the column vector of $h_{ij}(\hat{p})$'s, and \hat{P}_r and \hat{P}_c are the values of $\tilde{P}_r = \text{diag}(p_r) - p_r p_r'$ and $\tilde{P}_c = \text{diag}(p_c) - p_c p_c'$ respectively, for $p = \hat{p}$, where $p_r = (p_{1+}, \dots, p_{r-1,+})'$ and $p_c = (p_{+1}, \dots, p_{+,c-1})'$.

Rao and Scott (1981) have shown that under the null hypothesis $H_0 : \tilde{h}(p) = \underline{0}$, $X^2 \approx \sum_{i=1}^b \delta_{0i} W_i$, where δ_i 's are the eigenvalues of $\tilde{D}_h = (\tilde{P}_r^{-1} \otimes \tilde{P}_c^{-1})\tilde{V}_h$, $\delta_1 \geq \dots \geq \delta_b > 0$, W_1, \dots, W_b are independent χ_1^2 random variables and δ_{0i} is the value of δ_i under H_0 . Here \tilde{V}_h/n is the covariance matrix of $\tilde{h}(p)$ and its consistent estimator \hat{V}_h/n , the covariance matrix of $\tilde{h}(\hat{p})$, is obtained using the linearization method (see Fellegi, 1980) or the balanced repeated replication or the jackknife method. But we use the bootstrap method in our application. The modified statistic $X^2/\hat{\delta}$ is a χ_b^2 random variable, where the $\hat{\delta}_i$'s are the eigenvalues of $(\hat{P}_r^{-1} \otimes \hat{P}_c^{-1})\hat{V}_h$ and $\hat{\delta} = \sum_{i=1}^b \hat{\delta}_i/b$ with $b = (r-1)(c-1)$.

1.1.2 Brier (1980) Model

Brier formulated the Multinomial-Dirichlet distribution as a model for contingency tables generated by cluster sampling schemes. This model allows an arbitrary number of response categories and an arbitrary cluster size. In the model, given the cell probabilities indexed by the cluster indicators, the cell counts are assumed to have multinomial distribution. To accommodate the cluster effects, these cell probabilities are assigned the same Dirichlet distribution with independence over clusters.

Let $\underline{n}_i = (n_{i1}, \dots, n_{ik})$ be the vector of observed counts for the i^{th} ($i = 1, \dots, \ell$) cluster classified into k distinct categories such that $\sum_{j=1}^k n_{ij} = n_i$. Let $\underline{\pi}_i = (\pi_{i1}, \pi_{i2}, \dots, \pi_{ik})$ be the vector of cell probabilities for the i^{th} cluster with π_{ij} to be the probability of a unit in the i^{th} cluster being classified into the j^{th} category. Then, Brier model is

$$\underline{n}_i | \underline{\pi}_i \stackrel{ind}{\sim} \text{Multinomial}(n_i, \underline{\pi}_i),$$

$$\underline{\pi}_i | \underline{\mu}, \tau \stackrel{iid}{\sim} \text{Dirichlet}(\underline{\mu}\tau),$$

where $\underline{\mu}$ and τ are to be specified. Note that a model for a simple random sampling occurs in the limit as τ goes to infinity. The unconditional covariance matrix of \underline{n} under cluster sampling is a constant times the covariance matrix under simple random sampling; see Brier (1980). This constant is the design effect (which is the ratio of variance of the statistic under complex design to that of simple random sampling) and letting $n = \sum_{i=1}^{\ell} n_i$, it is $B = \frac{1}{n} \sum_{i=1}^{\ell} n_i \left(\frac{n_i + \tau}{1 + \tau} \right)$, a weighted average of $(n_i + \tau)/(1 + \tau)$, $i = 1, \dots, \ell$. Note here that this standard Multinomial-Dirichlet model provides the same design effect for the estimator of each cell probability of the two-way table (Brier, 1980). However, the second order Rao-Scott approximation provides different design effects.

Brier (1980) has shown that under the null hypothesis, the distributions of the Pearson chi-squared and likelihood ratio statistics (X^2 and G^2) are the multiples of chi-squared random variables. That is, the distribution is $B\chi_{k-s-1}^2$ as the number of clusters become large ($\rightarrow \infty$), where k denotes the number of distinct categories and s denotes the dimension of the parameter space under the null hypothesis. For example, in an $r \times c$ table with the

null hypothesis

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}, \quad i = 1, \dots, r, \quad j = 1, \dots, c,$$

we have $k = rc$ and $s = r + c - 2$ so that $k - s - 1 = (r - 1)(c - 1)$. After finding the consistent estimator, \hat{B} of B , a simple correction to the usual X^2 and G^2 is obtained as $X^2/\hat{B} \rightarrow \chi_{k-s-1}^2$ and $G^2/\hat{B} \rightarrow \chi_{k-s-1}^2$.

1.1.3 Geenens and Simar (2010, 2011)

Geenens and Simar (2010) address the problem of testing for independence between two categorical variables given a vector X of continuous covariates, focused on the case where X is a scalar continuous variable under simple random sampling. They proposed two nonparametric tests which generalize the chi-squared and the likelihood ratio tests. The procedure is based on a kernel estimator of the conditional probabilities.

Consider a sample of n individuals in a table cross-classified by two categorical variables R and S with r and s levels respectively. Let, $\pi_{ij} = P(R = i, S = j)$, $1 \leq i \leq r$, $1 \leq j \leq c$, be the probability that a given individual belongs to the cell (i, j) of the table. Then, the unconditional independence hypothesis is

$$H_0 : \pi_{ij} = \pi_{i.}\pi_{.j} \quad \forall(i, j),$$

where $\pi_{i.} = P(R = i) = \sum_{j=1}^c \pi_{ij}$ and $\pi_{.j} = P(S = j) = \sum_{i=1}^r \pi_{ij}$. The corresponding conditional independence hypothesis is

$$H_0 : \pi_{ij}(\chi) = \pi_{i.}(\chi)\pi_{.j}(\chi) \quad \forall \chi \in S_\chi, \forall(i, j),$$

where $S_\chi \subset R^p$ is the support of X and $\pi(\chi) = \{\pi_{ij}(\chi) : 1 \leq i \leq r, 1 \leq j \leq c\}$, the joint distribution of R and S conditional on X with $\pi_{ij}(\chi) = P(R = i, S = j | X = \chi)$.

The test procedure involves two steps. First, for any χ in S_χ , obtain a pointwise divergence criterion between the estimated joint conditional distribution of R and S given $X = \chi$ and the product of the marginal conditional distribution of R and S given $X = \chi$. The divergence criterion is basically the generalization of the classical chi-squared or the likelihood ratio criteria. The conditional distributions, $\pi(\chi) = E(Z | X = \chi)$, where $Z = (Z^{(11)}, Z^{(12)}, \dots, Z^{(rs)})'$

with $Z^{(ij)}$ taking the value 1 if the individual belongs to $(i, j)^{\text{th}}$ cell and 0 otherwise, regarded as regression functions, are nonparametrically estimated by Nadaraya-Watson-like estimators. Second, the pointwise divergence is integrated with respect to χ in order to evaluate this divergence on the whole support of S_χ which provides the test statistic.

Geenens and Simar (2011) considered the conditional joint distribution of two categorical variables given a set of explanatory variables X when analyzing a contingency table. They proposed a semiparametric model to estimate the conditional probabilities. In doing so, they assumed that the effect of the vector of covariates (X) on the cell probabilities can be captured by a single index $\theta_0^t X$, which is a linear combination of the initial covariates X . The estimation then involves two steps: first, estimate the coefficients (θ_0) of the linear combination and second, estimate the functions linking this index to the related conditional probabilities.

1.2 Surrogate Sampler, Nandram (2007)

Nandram (2007) used surrogate sampling to convert data obtained through a selection bias mechanism to provide an equivalent simple random sample. Nandram (2007) considered a problem in which a sample is drawn from a finite population but because of selection bias, the sample is not a random sample from the original finite population. In fact, the original sample is a random sample from a weighted distribution and one can convert this sample to a surrogate sample from the original distribution. This surrogate sample can be used to make an inference about the original finite population without any further consideration about the biased sample.

Let y_i , $i = 1, \dots, N$, denote the finite population values and let $p(\underline{y} | \theta_1)$ denote the probability distribution that describes the finite population (i.e., census). When a random sample is taken from this finite population, it is perturbed by the weight function $w(\underline{y}; \theta_1, \theta_2)$ to produce a sample from the new probability distribution $q(\underline{y} | \theta_1, \theta_2)$. That is, a representative sample is observed from

$$q(\underline{y} | \theta_1, \theta_2) = w(\underline{y}; \theta_1, \theta_2)p(\underline{y} | \theta_1).$$

The idea here is to create a surrogate (representative) sample from the original finite pop-

ulation using $p(y | \theta_1)$ and then make an inference about the finite population proportion. The Bayesian analysis is used to convert the biased sample into a random sample from the finite population. Nandram, Bhatta, Bhadra and Shen (2012) used the surrogate sampler to infer about a finite population proportion using data from a possibly biased sample.

For our Bayesian test of independence between two categorical variables, we use the Bayes factor as our test statistics. It is easier to calculate the Bayes factor under simple random sampling design because we have simple and closed form formula. This motivated us to convert the cluster sample into an equivalent simple random sample. We discuss the Bayes factor calculation for the simple random sample below.

1.3 Bayes Factor for a Test of Independence

For the $r \times c$ categorical table, we can consider two multinomial-Dirichlet models, one with association and the other with no association.

The model with association is

$$\underline{n} | \underline{\pi} \sim \text{Multinomial}(n, \underline{\pi}) \text{ and } \underline{\pi} \sim \text{Dirichlet}(\underline{u}), \quad (1.2)$$

where \underline{u} is specified.

Letting $\pi_{jk}^* = \pi_j^{(1)} \pi_k^{(2)}$, $j = 1, \dots, r$, $k = 1, \dots, c$, the model with no association is

$$\begin{aligned} \underline{n} | \underline{\pi}^{(1)}, \underline{\pi}^{(2)} &\sim \text{Multinomial}(n, \underline{\pi}^*), \\ \underline{\pi}^{(1)} &\sim \text{Dirichlet}(\underline{v}) \text{ and independently } \underline{\pi}^{(2)} \sim \text{Dirichlet}(\underline{w}), \end{aligned} \quad (1.3)$$

where $\underline{\pi}^{(1)}$ and $\underline{\pi}^{(2)}$ have r and c components respectively and \underline{v} and \underline{w} are specified.

Therefore, integrating out $\underline{\pi}^{(1)}$ and $\underline{\pi}^{(2)}$ from (1.3) and $\underline{\pi}$ from (1.2), it is easy to show that the marginal likelihood with association (as) is $p_{\text{as}}(\underline{n}) = \frac{n!}{\prod_{j=1}^r \prod_{k=1}^c n_{jk}!} \frac{D(\underline{n} + \underline{u})}{D(\underline{u})}$, and with no association (nas) is

$$p_{\text{nas}}(\underline{n}) = p_{\text{as}}(\underline{n}) \left\{ \frac{D(\underline{n}^{(1)} + \underline{v})}{D(\underline{v})} \frac{D(\underline{n}^{(2)} + \underline{w})}{D(\underline{w})} / \frac{D(\underline{n} + \underline{u})}{D(\underline{u})} \right\}, \quad (1.4)$$

where $\underline{n}^{(1)} = (n_{1.}, \dots, n_{r.})'$ and $\underline{n}^{(2)} = (n_{.1}, \dots, n_{.c})'$. Thus, using (1.4) the Bayes factor (BF)

is given by

$$BF = p_{\text{as}}(\underline{n})/p_{\text{nas}}(\underline{n}), \quad (1.5)$$

which provides evidence for association relative to no association. With Jeffreys' prior (i.e., elements of y , v and w are all 0.5) there is no simplification to (1.4) or (1.5).

However, for the special case where $y = \underline{1}$, $v = \underline{1}$ and $w = \underline{1}$ (i.e., uniform priors), we have $p_{\text{as}}(\underline{n}) = (rc - 1)!n!/(n + rc - 1)!$ and with no association (nas),

$$p_{\text{nas}}(\underline{n}) = p_{\text{as}}(\underline{n}) \frac{(r - 1)!(c - 1)!}{(rc - 1)!} \frac{(n + rc - 1)!}{(n + r - 1)!(n + c - 1)!} \frac{\prod_{j=1}^r n_{j\cdot}! \prod_{k=1}^c n_{\cdot k}!}{\prod_{j=1}^r \prod_{k=1}^c n_{jk}!}. \quad (1.6)$$

Looking for evidence of association, we use the rule of thumb of the log-Bayes factor, (0 - 1), not worth more than a bare mention; (1 - 3), positive; (3 - 5), strong; 5+, very strong; see Kass and Raftery (1995).

1.4 Applications

To illustrate our methodologies, we use two examples based on the Third International Mathematics and Science Study (1995) and the Trend in International Mathematics and Science Study (2007). The purpose of these studies is to examine students achievement in mathematics and science according to some variables in participating countries at high school level. In both studies, the data were collected from a stratified two-stage cluster sampling design.

1.4.1 Third International Mathematics and Science Study (TIMSS 1995)

We use the third grade population data consisting of 2477 students ¹ from United States. Here, the clusters are schools, and the units are the students. There are four strata: Northeast, South, Central and West regions of the US. We consider three of the variables in the survey: mathematics test scores (below average, average and above average), science test scores (below average, average and above average), and the communities the students come from (village or rural area, outskirts of a town or city and close to the center of a town or city).

¹(ftp://ftp.wiley.com/public/sci_tech_med/finite_population/)

Within each stratum, we study the association between mathematics test scores (MTS) and communities (COM), and science test scores (STS) and communities (COM), so there are eight examples.

1.4.2 Trend in International Mathematics and Science Study (TIMSS 2007)

We use data consisting of 7,896 fourth grade US students. It is an entire population data of participating schools. Here, the clusters are the schools and communities are the strata. There are six communities classified according to the size of the populations. We also have some observed student level and cluster level characteristics (or covariates) in the data. The student level covariates are (i) Sex, (ii) How often do you speak English at home?, (iii) Index of self confidence in learning math, (iv) Index of self confidence in learning science and (v) Race. The cluster level covariates are (i) Approximately what percentage of students in school come from economically affluent homes, (ii) Percent of free lunch-categorized and (iii) Total school enrollment in all grades. This is the public data available from the National Center for Educational Statistics ².

In this application, we consider two variables: mathematics test score (below average and above average) and science test scores (below average and above average). We are interested in studying the association between mathematics and science scores represented as categorical variables within each community, so there are six examples. The examination of the association between mathematics and science achievements is important because it may help mathematics and science educators assess the need for curriculum integration advocated by several professional organizations in the US and other nations.

To assess the intersubject relationship between the continuous mathematics and science scores TIMSS researchers have adopted modern assessment methodology. For this five plausible scores have been imputed in each subject area, and “one set of the imputed plausible scores can be considered as good as another” (Gonzalez and Smith, 1997, ch. 6, p. 3). Plausible values represent what the true performance of an individual might have been, had it been observed. This interchangeability also suggests equivalency of the design effects (deff) among

²(<http://nces.ed.gov/timss/>)

the plausible scores from stratified sample design (Wang, 2005). Then under an assumption of the invariant deff values between mathematics and science scores, the AM software is adopted to compute correlation coefficients between the variables. AM is a statistical software package developed by the American Institute of Research (AIR) for analyzing data from complex samples, especially large-scale assessments such as the National Assessment of Educational Progress (NAEP) and the Third International Mathematics and Science Studies (TIMSS). Wang (2005) examined the relationship between mathematics and science achievement based on student test scores using correlation coefficients.

1.5 Plan of Dissertation

The dissertation has three additional chapters.

In Chapter 2, we describe the test of independence without covariates. We have developed a method to overcome the limitations (e.g. small sample size) in the Rao-Scott methodology.

In Chapter 3, we describe the test of independence with covariates. We have found a related nonparametric conditional test under the simple random sampling design. However, we did not find any literature for complex survey data.

Finally, in Chapter 4 we summarize our contribution and present concluding remarks. We also discuss future research work that can be carried out within our framework. These can also improve our proposed methodology.

Chapter 2

A Test of Independence Without Covariates

In Chapter 2 we discuss the Bayesian test of independence in a two-way contingency table without covariates when the data are obtained from the two-stage cluster sampling with simple random sampling at both stages. We develop a hierarchical Bayesian model to convert the observed cluster data into simple random samples. This provides surrogate samples. We apply our methodology to the TIMSS 1995 data and present the results.

For many large complex surveys the Rao-Scott corrections to the standard chi-squared (or likelihood ratio) statistic provide appropriate inference. For smaller surveys, however, the Rao-Scott corrections may not be accurate, partly because the chi-squared test is inaccurate. We present a method to overcome the limitations in the Rao-Scott methodology.

In two-stage cluster sampling, a simple random sample of ℓ clusters (primary sampling unit or psu's) is selected, and within the i^{th} sampled cluster a simple random sample of n_i units (secondary sampling units or ssu's) is selected. Note here that data are obtained from a clustered super population in which each unit has exactly one of S characteristics ($S = rc$ for a $r \times c$ categorical table). Let n_{ijk} denote the counts in the $(j, k)^{\text{th}}$ cell of the $r \times c$ table constructed from the i^{th} cluster; we call this table the i^{th} cluster table. Analogously, let $n_{jk} = \sum_{i=1}^{\ell} n_{ijk}$ be the cell counts for the $(j, k)^{\text{th}}$ cell of the table of total counts. We will call the table of total counts the *total table*. We are interest in the test of independence of two categorical variables in the $r \times c$ total table.

The idea in our method is to simulate a large sample of total tables under simple random sampling, and compute the Bayes factor for a test of independence from each simulated

table. Then, the distribution of Bayes factors is obtained which is used for a final test of independence. While we present the methodology for two-stage cluster sampling, a special case, the approach can be extended to other complex surveys. Under our method, first, we start with a model appropriate for simple random sampling and elaborate it to accommodate the more complex sample design. Next, fitting the observed data to the model for complex sample design, we make inference for the (population) parameter, θ , of the initial model (i.e., draw M' samples from the posterior distribution of θ). Then we use $\theta^{(1)}, \dots, \theta^{(M')}$ to draw simple random samples consistent with the observed data. The data from these simple random samples are then used to make the required inferences (e.g., to test independence in a contingency table using a Bayes factor).

2.1 Hierarchical Bayesian Model

We string out the counts in the total table to an array of $S = rc$ cells (i.e., n_s , $s = 1, \dots, S$). If we assume simple random sampling, our Bayesian model is

$$\begin{aligned} \underline{n} \mid \tilde{\pi} &\sim \text{Multinomial}(n, \tilde{\pi}), \\ \tilde{\pi} &\sim \text{Dirichlet}(\underline{1}), \end{aligned} \tag{2.1}$$

where $\underline{n} = (n_1, \dots, n_S)$, $\tilde{\pi} = (\tilde{\pi}_1, \dots, \tilde{\pi}_S)$, $n = \sum_{s=1}^S n_s$ and $\underline{1}$ is a vector of S ones. We call this model with simple random sampling MSRS. Typically, the total table will have large counts relative to the cluster tables, so that the uniform prior is approximately noninformative (i.e., posterior mode is the same as the maximum likelihood estimator). It is possible to have a few cells with zero counts, but most of the cell counts are expected to be relatively large.

We take care of the clustering by assuming that

$$\underline{n}_i \mid \underline{a}_i \stackrel{\text{ind}}{\sim} \text{Multinomial}(n_i, \underline{a}_i), \tag{2.2}$$

where $\underline{n}_i = (n_{i1}, \dots, n_{iS})$, $n_i = \sum_{s=1}^S n_{is}$ and $a_{is} = \alpha_{is}\pi_s$, $i = 1, \dots, \ell$, $s = 1, \dots, S$. Note here that $\alpha_{is}\pi_s$ is the probability that a unit has the s^{th} characteristic within the i^{th} cluster of the super population and π_s , $s = 1, \dots, S$, are the probabilities corresponding to a homogeneous superpopulation (i.e., there are no clusters). In (2.2) we have the constraints $\{\sum_{s=1}^S \alpha_{is}\pi_s =$

1, $i = 1, \dots, \ell, \sum_{s=1}^S \pi_s = 1, \alpha_{is} > 0, \pi_s > 0$ }. Here, the α_{is} are used to adjust for the clustering.

A priori we take,

$$\alpha_{is} \mid \tau_s, \nu \stackrel{ind}{\sim} \text{Gamma}(\tau_s, \tau_s \nu), s = 1, \dots, S \quad (2.3)$$

and

$$\underline{\pi} \sim \text{Dirichlet}(\underline{1}), \quad (2.4)$$

where $\underline{\pi} = (\pi_1, \dots, \pi_S)$ and $\tau_s, s = 1, \dots, S$, are to be specified. Noting that in $a_{is} = \alpha_{is}\pi_s$, neither the α_{is} nor the π_s are identifiable. This is true because while the number of cells in the i^{th} cluster table is S , the number of parameters corresponding to the i^{th} cluster is $2(S - 1)$. Thus, we choose to specify the τ_s to allow both α_{is} and π_s to be identifiable.

We note two important features of this model. First, a model for simple random sampling is a special case of ours. This is easily seen by setting $\alpha_{is} \equiv 1$. Second, by construction, the model gives a positive correlation among the units in a cluster and this correlation varies with the cell of the contingency table. To show this, let $I_{isj} = 1$ if a j^{th} secondary sampling unit (ssu) falls in the s^{th} cell and $I_{isj} = 0$ otherwise. Then, given α_{is} and π_s , $I_{isj} \stackrel{iid}{\sim} \text{Bernoulli}(\alpha_{is}\pi_s)$. After some algebraic manipulation, it follows that $\text{var}(I_{isj}) = (\nu S - 1)/\nu^2 S^2$, $\nu \geq S^{-1}$, independent of s , and $\text{cov}(I_{isj}, I_{isj'}) = \{2\tau_s^{-1} + (S - 1)/S\}/S(S + 1)\nu^2$, $j \neq j'$, positive, not independent of s . Therefore, $\text{cor}(I_{isj}, I_{isj'}) = \{S(2\tau_s^{-1} + 1) - 1\}/(S + 1)(\nu S - 1)$, $j \neq j'$, and by the Cauchy-Schwarz inequality the intracluster correlation lies in $(0, 1)$ provided that $\nu > S^{-1}$. Because the correlation varies with the cell of the contingency table, we have different design effects for the estimators of the cell probabilities of the total table. Henceforth, we let $\nu_o = S^{-1}$ so that $\nu > \nu_o$.

Finally, for ν , we assume a standard noninformative prior,

$$p(\nu) \propto 1/\nu, \nu > 0. \quad (2.5)$$

Note that the joint prior density of the α_{is} , π_s and ν must satisfy the above constraints, $\{\sum_{s=1}^S \alpha_{is}\pi_s = 1, i = 1, \dots, \ell, \sum_{s=1}^S \pi_s = 1, \alpha_{is} > 0, \pi_s > 0\}$. This is our model for a two-stage cluster sampling design and we will call it MCSD.

It is easy to fit MSRS. In fact, under MSRS,

$$\tilde{\pi} \mid \underline{n} \sim \text{Dirichlet}(\underline{n} + \underline{1}).$$

However, the Bayesian model under cluster sampling is much more complex partly because of the constraints and the awkwardness of the a_{is} . The joint posterior density is obtained in Appendix A. Because the prior density in (A.2) is improper, the joint posterior density in (A.4) may be improper. We next show that the joint posterior density in (A.4) is proper.

Theorem 2.1.1. *The joint posterior density in (A.4) is proper.*

Proof: We make the transformation $t_{is} = \alpha_{is}\pi_s$, $s = 1, \dots, S-1$, $i = 1, \dots, \ell$, keeping the π_s untransformed. The Jacobian of the transformation is $\left(\prod_{s=1}^{S-1} \pi_s\right)^{-\ell}$ and the joint posterior density becomes

$$\begin{aligned} p(\underline{t}, \underline{\pi}_{(S)}, \nu \mid \underline{n}) &\propto \\ \nu^{\ell b-1} e^{-\nu \sum_{i=1}^{\ell} \left(\sum_{s=1}^{S-1} \tau_s \frac{t_{is}}{\pi_s} + \tau_S \frac{1 - \sum_{s=1}^{S-1} t_{is}}{1 - \sum_{s=1}^{S-1} \pi_s} \right)} &\prod_{i=1}^{\ell} \left[\left(\prod_{s=1}^{S-1} t_{is}^{n_{is} + \tau_s - 1} \right) \left(1 - \sum_{s=1}^{S-1} t_{is} \right)^{n_{iS} + \tau_S - 1} \right. \\ &\left. \times \left\{ \left(\prod_{s=1}^{S-1} \pi_s^{\tau_s} \right) \left(1 - \sum_{s=1}^{S-1} \pi_s \right)^{\tau_S} \right\}^{-1} \right], (\underline{t}, \underline{\pi}_{(S)}, \nu) \in T^*, \end{aligned}$$

where

$$T^* = \left\{ (\underline{t}, \underline{\pi}_{(S)}, \nu) : 0 < \sum_{s=1}^{S-1} t_{is}, \sum_{s=1}^S \pi_s < 1, t_{is}, \pi_s > 0, i = 1, \dots, \ell, s = 1, \dots, S-1, \nu > \nu_o \right\}.$$

Now, assuming $\ell b > 1$ and letting $F_{\ell b}(a) = \int_0^a t^{\ell b-1} e^{-t} / \Gamma(\ell b) dt$, the cdf of a gamma random variable and integrating out ν , we get

$$\begin{aligned} p(\underline{t}, \underline{\pi}_{(S)} \mid \underline{n}) &\propto \{1 - F_{\ell b}(A\nu_o)\} A^{-\ell b} \\ &\times \prod_{i=1}^{\ell} \left[\left(\prod_{s=1}^{S-1} t_{is}^{n_{is} + \tau_s - 1} \right) \left(1 - \sum_{s=1}^{S-1} t_{is} \right)^{n_{iS} + \tau_S - 1} \left\{ \left(\prod_{s=1}^{S-1} \pi_s^{\tau_s} \right) \left(1 - \sum_{s=1}^{S-1} \pi_s \right)^{\tau_S} \right\}^{-1} \right], (\underline{t}, \underline{\pi}_{(S)}) \in \tilde{T}^*, \end{aligned} \tag{2.6}$$

where

$$\tilde{T}^* = \left\{ (\underline{t}, \underline{\pi}) : 0 < \sum_{s=1}^{S-1} t_{is}, \sum_{s=1}^S \pi_s < 1, t_{is}, \pi_s > 0, i = 1, \dots, \ell, s = 1, \dots, S-1 \right\},$$

and

$$A = \sum_{i=1}^{\ell} \left\{ \sum_{s=1}^{S-1} \tau_s \frac{t_{is}}{\pi_s} + \tau_S \left(\frac{1 - \sum_{s=1}^{S-1} t_{is}}{1 - \sum_{s=1}^{S-1} \pi_s} \right) \right\}.$$

Since $p(\underline{t}, \underline{\pi}_{(S)} \mid \underline{n})$ is finite on any compact subset of \tilde{T}^* , the integral of $p(\underline{t}, \underline{\pi}_{(S)} \mid \underline{n})$ over any compact subset of \tilde{T}^* is finite. Thus, the joint posterior density $p(\underline{t}, \underline{\pi}_{(S)}, \nu \mid \underline{n})$ is proper.

2.2 Computations, Bayes Factor and Specifications

Letting \underline{n} denote the observed data from the total table and \underline{n} the vector of surrogate sample counts for the total table, we need to generate sample from

$$f_{SRS}(\hat{\underline{n}} \mid \underline{n}) = \int f_{SRS}(\hat{\underline{n}} \mid \underline{\pi}, \underline{n}) f_{CL}(\underline{\pi} \mid \underline{n}) d\underline{\pi}. \quad (2.7)$$

In (2.7) f_{SRS} indicates that $\hat{\underline{n}}$ are the surrogate cell counts appropriate to simple random sampling and f_{CL} is the posterior density of $\underline{\pi}$ using the model for the observed cluster data (MCSD). In Section 2.2.1 we show how to generate samples from $f_{CL}(\underline{\pi} \mid \underline{n})$ using (2.6). In section 2.2.2 we show how to obtain simple random sample from $f_{SRS}(\hat{\underline{n}} \mid \underline{\pi}, \underline{n})$ using (2.1). We then show how to obtain the Bayes factor, and we also show how to specify the parameters $\tau_s, s = 1, \dots, S$. We use a computational method which ensures that our method is more accurate and at least as fast as the methods of Rao and Scott (1981).

2.2.1 Computations

As is apparent, the joint posterior density is complicated, and so we need a sampling based method to draw samples from it. We obtain random draws from an approximation of the joint posterior density and then use the sampling importance resampling (SIR) algorithm (Gelman, Carlin, Stern and Rubin 2004, Ch. 12) to subsample these draws to obtain samples from $\underline{\pi} \mid \underline{n}$; this gives us the required samples of $\underline{\pi}$. Note that we are not using Markov chain Monte Carlo methods because we want to avoid monitoring which will not make our algorithm as fast the Rao-Scott methods.

Using a heuristic argument we conjecture that an approximation which satisfies four properties may be useful. First, an approximation should have some dependence between the t_{is} and the π_s ; see (2.6). Second, t_{is} and π_s should have similar forms. Third, the distributions of t_{is} and π_s should be functions of the data (i.e., the cell counts of the cluster tables) to allow the data to have direct influence on these distributions. Fourth, the computations of the approximation must be fast and require no monitoring. To approximate the joint density of \underline{t} and $\underline{\pi}$, we take $\underline{t}_i = (t_{i1}, \dots, t_{iS})$, $i = 1, \dots, \ell$, given $\underline{\pi}$ and \underline{n} to be independent, giving

$$p_a(\underline{t}, \underline{\pi} \mid \underline{n}) = \left\{ \prod_{i=1}^{\ell} p_a(\underline{t}_i \mid \underline{\pi}, \underline{n}) \right\} p_a(\underline{\pi} \mid \underline{n}), \quad (2.8)$$

where $p_a(\underline{t}_i \mid \underline{\pi}, \underline{n})$ and $p_a(\underline{\pi} \mid \underline{n})$ are determined next.

First, to obtain the approximation, $p_a(\underline{\pi} \mid \underline{n})$, we consider the posterior density under simple random sampling. Here,

$$p^*(\underline{\pi} \mid \underline{n}) \propto \prod_{s=1}^S \pi_s^{n_{\cdot s} + 1}, \quad \sum_{s=1}^S \pi_s = 1.$$

Our intuition is that the correct posterior density under clustering sampling should be related to this posterior density under simple random sampling. However, it should reflect the clustering through the design effects. Thus, we make two additional adjustments to $p^*(\underline{\pi} \mid \underline{n})$. First, by penalizing $n_{\cdot s}$, $s = 1, \dots, S$, we replace $n_{\cdot s}$ by $n_{\cdot s}/\delta_s$ where δ_s are design effects, possibly all the same as in Brier's method. Second, to make this dependent on τ_s (suggested by the term in π_s in (2.6)), we add τ_s to $n_{\cdot s}/\delta_s + 1$ to get the approximate posterior density, $p_a(\underline{\pi} \mid \underline{n})$,

$$\underline{\pi} \mid \underline{n} \sim \text{Dirichlet}(d), \quad (2.9)$$

where $d_s = n_{\cdot s}/\delta_s + \tau_s + 1$, $s = 1, \dots, S$ and τ_s , $s = 1, \dots, S$ are specified in Section 2.2.3.

Second, note that ignoring the term $(1 - F(A\nu_o))A^{-\ell b}$ and the constraints, the conditional posterior density in (2.6) is of the form,

$$p^{**}(\underline{t} \mid \underline{\pi}, \underline{n}) \propto \prod_{i=1}^{\ell} \prod_{s=1}^S t_{is}^{n_{is} + \tau_s - 1}, \quad t_{is} > 0, \quad s = 1, \dots, S, \quad \sum_{s=1}^S t_{is} = 1, \quad i = 1, \dots, \ell.$$

That is, approximately $\underline{t}_i \mid \underline{n} \stackrel{ind}{\sim} \text{Dirichlet}(n_i + \underline{\tau})$. We allow this to be dependent on $\underline{\pi}$ by replacing n_{is} with $n_i \pi_s$. Then approximately $\underline{t}_i \mid \underline{\pi}, \underline{n} \stackrel{ind}{\sim} \text{Dirichlet}(n_i \underline{\pi}_s + \underline{\tau})$. Adding unity

to the Dirichlet parameters to increase computational stability, the final approximation, $p_a(\underline{t}_i | \underline{\pi}, \underline{n})$ of the conditional posterior distribution of $\underline{t}_i | \underline{\pi}, \underline{n}$ is

$$\underline{t}_i | \underline{\pi}, \underline{n} \stackrel{ind}{\sim} \text{Dirichlet}(b_i), \quad (2.10)$$

where $b_{is} = n_i \pi_s + \tau_s + 1, i = 1, \dots, \ell, s = 1, \dots, S$. Observe that (2.9) and (2.10) have similar forms.

We now show how to carry out the SIR algorithm. To obtain the probability of selecting each sampled iterate, we need to study the ratio,

$$R(\underline{t}, \underline{\pi}) = \frac{p(\underline{t}, \underline{\pi} | \underline{n})}{p_a(\underline{t}, \underline{\pi} | \underline{n})},$$

where $p(\underline{t}, \underline{\pi} | \underline{n})$ and $p_a(\underline{t}, \underline{\pi} | \underline{n})$ are given, respectively, in (2.6) and (2.8). Simplifying, we get,

$$R(\underline{t}, \underline{\pi}) = C \frac{\{1 - F_{lb}(A\nu_o)\} \prod_{i=1}^{\ell} [\{\prod_{s=1}^S t_{is}^{n_{is} - n_i \pi_s - 1}\} D(n_i \underline{\pi} + \underline{\tau} + \underline{1})]}{\{\prod_{s=1}^S \pi_s^{n_s / \delta_s + (\ell+1)\tau_s}\} A^{b\ell}}, \quad (2.11)$$

where strictly $0 < \pi_s < 1, 0 < t_{is} < 1, D(\cdot)$ is the Dirichlet function and C is a proportionality constant. Note that, by construction, $R(\underline{t}, \underline{\pi})$ is bounded because both $p(\underline{t}, \underline{\pi} | \underline{n})$ and $p_a(\underline{t}, \underline{\pi} | \underline{n})$ are bounded. The SIR algorithm requires $R(\underline{t}, \underline{\pi})$ to be bounded.

We use 10% subsampling. We draw $\tilde{M} = 10,000$ samples from the approximate joint posterior density in (2.8). This is obtained using the composition rule by first drawing $\underline{\pi}$ from (2.9) and, in turn, drawing \underline{t}_i from (2.10). Letting $\Omega^{(h)} = (\underline{t}^{(h)}, \underline{\pi}^{(h)}), h = 1, \dots, \tilde{M}$, the subsampling probabilities are $W_h = R(\Omega^{(h)}) / \sum_{h'=1}^{\tilde{M}} R(\Omega^{(h')})$, $h = 1, \dots, \tilde{M}$, where $R(\cdot)$ is given in (2.11). Then we sample 10% of the \tilde{M} samples without replacement to get $M = .10\tilde{M}$ samples (drawing without replacement is a sensible procedure to avoid some values being sampled repeatedly). Thus, we finally have samples from the posterior density of $\underline{\pi}$.

2.2.2 Bayes Factor

Having obtained samples from the posterior density of $\underline{\pi}$, we can now obtain samples from the distribution of the Bayes factor. Let $\underline{\pi}^{(h)}, h = 1, \dots, M$, denote the M samples from our MCSD (i.e., cluster model). Then, we draw $\hat{\underline{n}}^{(h)}$ from the total table,

$$\hat{\underline{n}}^{(h)} \stackrel{ind}{\sim} \text{Multinomial}\{n, \underline{\pi}^{(h)}\}, h = 1, \dots, M.$$

Here, $\hat{n}^{(h)}$ is surrogate data because the original total table (observed data) has now been converted and a model for simple random sampling is appropriate. Thus, we have M surrogates for the total table. Now, to compute a sample of M values of the Bayes factor, we fit a model of association and a model of no association to the surrogate data, $\hat{n}^{(h)}$, $h = 1, \dots, M$, each surrogate in turn. We take the model of association to be

$$\underline{n}^{(h)} \sim \text{Multinomial}(n, \underline{\pi}), \quad \underline{\pi} \sim \text{Dirichlet}(\underline{u}), h = 1, \dots, M, \quad (2.12)$$

where $u_s = .5$, $s = 1, \dots, S$, for Jeffreys' prior (proper prior). Letting $\pi_{jk}^* = \pi_j^{(1)}\pi_k^{(2)}$, $j = 1, \dots, r$, $k = 1, \dots, c$, the model with no association is

$$\underline{n}^{(h)} \mid \underline{\pi}_\sim^{(1)}, \underline{\pi}_\sim^{(2)} \sim \text{Multinomial}(n, \underline{\pi}_\sim^*),$$

$$\underline{\pi}_\sim^{(1)} \sim \text{Dirichlet}(\underline{v}) \text{ and independently } \underline{\pi}_\sim^{(2)} \sim \text{Dirichlet}(\underline{w}), \quad (2.13)$$

where $v_j = .5$, $j = 1, \dots, r$ and $w_k = .5$, $k = 1, \dots, c$. It is worth noting that, while the computation of Bayes factor requires proper prior distributions, proper priors are not required in MCSD as long as the posterior density (2.6) is proper (as we have shown in Appendix A). However, we do need proper priors in (2.12). Inference should not be sensitive to moderate departures from Jeffreys' prior because the cell counts of the total table are expected to be relatively large.

In Section (1.3) we present the Bayes factor for a test of independence for the total table which is given by

$$BF^{(h)} = p_{\text{nas}}(\underline{n}^{(h)})/p_{\text{as}}(\underline{n}^{(h)}), h = 1, \dots, M,$$

where $p_{\text{nas}}(\underline{n}^{(h)})$ and $p_{\text{as}}(\underline{n}^{(h)})$ are, respectively, the marginal likelihoods under the models with no association (nas) and association (as). In Appendix C, we show how to obtain the mode of the posterior distribution of the Bayes factor. It is straightforward to obtain other summaries of the Bayes factor.

Thus, our method obtains M estimates of the Bayes factor and these estimates, in turn, provide an estimate of the empirical distribution of the true Bayes factor. Our computations show that the entire procedure to obtain the M estimates of the Bayes factor and its distribution takes less than five seconds for data from small two-stage cluster sampling designs.

Henceforth, we will mostly work with the log-Bayes factor.

2.2.3 Specifications

We show how to specify the design effects δ_s and the τ_s . Note that while the τ_s are part of MSCD, δ_s are not part of MCS D and the δ_s only affect the computations.

We state and prove an important lemma about the maximum likelihood estimator (MLE) of the parameters of a gamma distribution in Appendix B. We will use this lemma repeatedly to specify the hyperparameters and the tuning constants.

Let $n'_{is}, i = 1, \dots, \ell, s = 1, \dots, S$, denote past data or data from a similar survey. We obtain estimates of α_{is} from the cluster tables with cell counts $n'_{is}, i = 1, \dots, \ell, s = 1, \dots, S$, adding 0.5 because of some zero cell counts. First, define $\hat{p}_{is} = (n'_{is} + .5)/(n'_i + .5S)$, $\hat{\pi}_s = (n'_s + .5)/(n' + .5S)$ and $\hat{\alpha}_{is} = \hat{p}_{is}/\hat{\pi}_s, i = 1, \dots, \ell, s = 1, \dots, S$. We use this form for the $\hat{\alpha}_{is}$ because under (2.2) only, $E(n'_{is}/n'_i) = \alpha_{is}\pi_s, i = 1, \dots, \ell, s = 1, \dots, S$. Therefore, removing the expectation on the left-hand side, we get $\hat{p}_{is} \approx \hat{\alpha}_{is}\hat{\pi}_s$. Then, we take

$$\hat{\alpha}_{is} \stackrel{iid}{\sim} \text{Gamma}(\tau_s, \tau_s\nu)$$

as in (2.3). First, pretending that the τ_s are equal and letting A denote the arithmetic mean of the $\hat{\alpha}_{is}$, the MLE of ν is $\hat{\nu} = A^{-1}$ as in Appendix B. Then, for τ_s a ‘profile’ log-likelihood is obtained by replacing ν in the log-likelihood function by A^{-1} . For each τ_s with ν fixed at A^{-1} , we obtain the MLE of τ_s by maximizing the profile log-likelihood function,

$$\tau_s \ln(\tau_s) - \tau_s \ln(A) + (\tau_s - 1) \ln(G_s) - \tau_s A_s/A - \ln \Gamma(\tau_s), s = 1, \dots, S,$$

where A_s and G_s are the arithmetic and geometric means of $\hat{\alpha}_{is}$. By an argument similar to Appendix B, the MLE exists and is unique. We use the Nelder-Mead algorithm to do the maximization.

We now show how to obtain the design effects for the computation. We consider the following simpler model for cluster sampling,

$$n_i | \underline{\pi}_i \stackrel{ind}{\sim} \text{Multinomial}(n_i, \underline{\pi}_i) \text{ and } \underline{\pi}_i \stackrel{iid}{\sim} \text{Dirichlet}(\underline{\mu}\phi),$$

where $n_i = \sum_{s=1}^S n_{is}$ is the number of ssu’s in the i^{th} cluster, $\underline{\pi}_i = (\pi_{i1}, \dots, \pi_{iS})$ and ϕ are

to be specified. Note that simple random sampling occurs in the limit as ϕ goes to infinity. The covariance matrix of \underline{n} under cluster sampling is a constant times the covariance matrix under simple random sampling; see Brier (1980). This constant is the design effect and, letting $n = \sum_{i=1}^{\ell} n_i$, it is $\frac{1}{n} \sum_{i=1}^{\ell} n_i(n_i + \phi)/(1 + \phi)$, a weighted average of $(n_i + \phi)/(1 + \phi)$, $i = 1, \dots, \ell$.

To specify τ , we start by using a method of moments estimator for $\underline{\mu}$ (i.e., $\hat{\mu}_s = \sum_{i=1}^{\ell} n_{is}/n$, $s = 1, \dots, S$). These are reasonably efficient estimators because they are formed from the total table. We obtain ϕ by maximizing the profile log-likelihood of the multinomial-Dirichlet model,

$$\sum_{i=1}^{\ell} \left[\sum_{s=1}^S \{ \ln \Gamma(n_{is} + \hat{\mu}_s \phi) - \ln \Gamma(\hat{\mu}_s \phi) \} - \{ \ln \Gamma(n_i + \phi) - \ln \Gamma(\phi) \} \right]$$

over $\phi > 0$. We denote the MLE of ϕ by $\hat{\phi}$ and it is easily obtained using the Nelder-Mead algorithm. Thus we take $\delta_s = \frac{1}{n} \sum_{i=1}^{\ell} n_i \left(\frac{n_i + \hat{\phi}}{1 + \hat{\phi}} \right)$, $s = 1, \dots, S$, equal.

2.3 Numerical Analysis

We discuss an illustrative example. This example suggests certain features which are investigated further in a simulation study.

We use the mode of the distribution of the Bayes factors, obtained from the surrogate total tables, for testing and the interquartile range of these Bayes factors for gauging this evidence. We interpret the mode using the rule of thumb of Kass and Raftery (1995). However, we share the philosophy that evidence cannot be measured by a single test and other tests (e.g., Rao-Scott test) should be used. It is not sensible to look at a single p-value or just the mode of the distribution of the Bayes factor.

2.3.1 Illustrative Example

To illustrate our methodology, we use data from the Third International Mathematics and Science Study (TIMSS). The data consist of 2477 students¹. Here, the clusters are schools while the units are the students. There are four strata: Northeast, South, Central and West regions of the US. We consider three of the variables in the survey: mathematics test scores

¹(ftp://ftp.wiley.com/public/sci_tech_med/finite_population/)

(below average, average and above average), science test scores (below average, average, above average) and the communities the students come from (village or rural area, outskirts of a town or city and close to the center of a town or city). Within each stratum, we study the association between mathematics test scores (MTS) and communities (COM) and science test scores (STS) and communities (COM), so there are eight examples. We assume that the finite population is a sample from a superpopulation.

In Table 2.1 we present the total tables for the eight examples (E1-E4 for MTS versus COM and E5-E8 for STS versus COM in each of the four regions). The number (ℓ) of clusters changes considerably over regions as does the number of observations. The intra-class correlations are moderately large and they change considerably over examples. The design effects (DEFs), obtained from Brier's model, are considerably larger than one. Thus, in all the examples, the cluster effect is substantial. Some of the observed counts in the total tables do not exceed 5. This is noticeable in cell (1,3) (below average in a town or city) in all examples except E3 and E6. In E4, cell (1,3) is 0 so standard X^2 and G^2 tests are not really applicable. In fact, for E4 the Rao-Scott first order test cannot be computed using SAS because it uses linearization or the jackknife to estimate the covariance matrix. We are able to compute the Rao-Scott test because we use the bootstrap method. Rao-Scott methods do not provide a sensible adjustment because in our case they correct X^2 and G^2 only for clustering, not for tables with small cell counts.

The 'posterior' design effects for the individual cells are presented in Table 2.2. These are different from those in Brier's method (see Section 2.2.3) and are computed as the diagonals of the posterior variance of π under the hierarchical Bayesian model specified by (2.2)-(2.5) and the posterior variance under the model for simple random sampling specified by (2.1). These design effects are considerably larger than 1. The average design effects, 9.01, 6.75, 5.91, 7.80, 8.34, 5.67, 6.71, 7.10 for E1-E8, are very similar to the design effects obtained from Brier's method given in Table 2.1. But what is more important is that the DEFs vary quite a bit over the cells for all examples except E3 and E6. Thus, Brier's method is inappropriate except, perhaps, for E3 and E6. With such large variations in DEFs across the cells the, Rao-Scott approximations are not expected to work well. This is particularly

true in E4 in which cell (1, 3) has a design effect of 25.65 corresponding to the zero count. For completeness we have also calculated the effective sample size (ESS), the sum of the ratios of the original cell counts of the total table divided by the corresponding design effects. As can be seen in the last row of Table 2.2, these are considerably smaller than the original sample size (see Table 2.1).

In Table 2.3, we present summaries of the Bayes factor obtained from our model. Again, our rule is the one described by Kass and Raftery (1995) applied to the mode of the distribution of the Bayes factor. For example, in example E1 the mode is 5.7 and according to Kass and Raftery (1995) there is ‘strong’ evidence against independence. For comparison, we also present the p-values obtained from the standard chi-squared test and Rao-Scott first order (RSF) and second order (RSS) approximations.

In example E1, RSF and RSS do not reject independence, while the chi-squared test and Bayes factor test show evidence against independence. The very strong evidence against independence shown by the chi-squared test may be due to ignoring the large cluster effect ($\rho = .56$, see Table 2.1). Except perhaps for E2 and E6, the Bayes factors show that there is some evidence for a strong dependence between mathematics test scores and community and science test scores and community. It is interesting that the tests based on chi-squared, RSF, RSS and Bayes factor agree in all examples except E1.

We also obtained the proportion, P , of estimated Bayes factors in the 1000 runs that are larger than the observed Bayes factor under the (incorrect) simple random sampling in the observed total table. If the cluster sampling design was a simple random sampling design, it seems reasonable that these Bayes factors should have a distribution symmetric around the observed Bayes factor obtained from simple random sampling. Thus, under simple random sampling these P s should be around .5. These are shown in the penultimate column of Table 2.3. However, these P s are significantly larger than .5, showing that the clustering effect our model accounts for is substantial.

Because Bayesian estimation procedures are much less sensitive to prior specifications than Bayesian hypothesis we have considered an estimation procedure as well. Based on the hierarchical Bayesian model we have obtained 95% credible intervals of the ratios,

$\pi_{jk}/p_jq_k, j = 1, \dots, r, k = 1, \dots, c$ where $p_j = \sum_{k=1}^c \pi_{jk}$ and $q_k = \sum_{j=1}^r \pi_{jk}$. Note that there are $S = rc = 9$ credible intervals. Then, we have computed the number, N , of 95% credible intervals of π_{jk}/p_jq_k containing 1 (e.g., see Nandram and Choi 2007 for a similar procedure). If some of these intervals do not contain 1, this provides some evidence against independence. The values of N , presented in the last column of Table 2.3, show some evidence of independence in examples E3, E5 and E8. Of course, these intervals are much too wide for this latter procedure to be particularly useful. Nevertheless it is sensible to consider it as well.

In Figure 2.1, we present the distributions of the Bayes factors obtained from the 1000 estimates of the Bayes factor from each of the eight examples. Looking at where most of the distribution lies, it shows that in E2 and E6 there is little evidence against independence and in the other examples there is much stronger evidence against independence. Note that calculating the distribution provides substantially more information than quoting a single summary but it is not done in practice.

In Table 3.7, we study the issue of sensitivity of the Bayes factor to the specification of $\tau_s, s = 1, \dots, S$ ($S = 9$). We set $\tau_s = \eta \hat{\tau}_s$ where we take $\eta = .5, 1, 2$ and $\hat{\tau}_s$ are the maximum likelihood estimates. The mode, median, the first and third quartiles and P all decrease as η changes from 0.5 to 2. However, the evidence against independence does not change markedly. This is true in all eight examples. We have also looked at sensitivity to the specification of the uniform prior for the model based on simple random sampling applied to the surrogate total tables. Small variations in Jeffreys prior show very small changes in the Bayes factor (e.g., changing .5 in Jeffreys prior to .10 or 1).

2.3.2 Simulation Study

We have performed a small simulation study to help understand these tests further. We consider three factors: dependence between the two categorical variables (weak, strong), the table density of the cluster tables (low, medium) and intracluster correlation (very small, small, moderate). The density of a total table is the total number of observations divided by the product of the number of clusters and the number of cells ($S = 9$ for a 3×3 table). We

have set the number of clusters at $\ell = 35$ and the table density, Δ , at 2 and 4 giving a total number of observations of 630 and 1260.

Corresponding to the nine cells (3×3 categorical table), let $\psi_s = 1, s = 2, 3, 4, 6, 7, 8$, (off-diagonal cells) and $\psi_s = \text{ind}, s = 1, 5, 9$ (diagonal cells) where ‘ind’ is to be specified. The cell probabilities are $\psi_s / \sum_{s=1}^S \psi_s, s = 1, \dots, S$. When the ψ_s are roughly the same (ind=1), there will be independence and when the diagonal ψ_s are larger than 1, there will be dependence (ind=2). For a 3×3 table with large cell counts, if the diagonal probabilities are twice the off diagonals, there will be strong dependence (ind=2). With an intracluster correlation of ρ , we set $\alpha_s = \{(1 - \rho)/\rho\} \psi_s / \sum_{s=1}^S \psi_s, s = 1, \dots, S$. For $i = 1, \dots, \ell$ we generate $\pi_i \stackrel{iid}{\sim}$ Dirichlet(α) to get the cell probabilities for the 35 cluster tables. We divide the total number of observations into the clusters with sizes, $n_i, i = 1, \dots, \ell$, based on multinomial distributions with equal cell probabilities. Finally, the cluster tables are generated independently from multinomial distributions with total counts n_i and cell probabilities π_i . We choose $\rho = .01, .10, .30$.

Thus, there are twelve ($2 \times 2 \times 3$) design points, and 100 cluster samples are generated at each design point. We perform our computations exactly as for the Third Grade population and obtain both the p-values and the Bayes factors from our model. We ‘average’ various quantities over the 100 replications at each design point. For example, in Table 2.5 the mode is the average of the 100 modes.

In Table 2.5, we present numerical summaries from the simulation study. As expected, the p-value of RSS is at least as large as the p-value from the chi-squared test and as ρ increases the design effects increase for all design points. It is good that the chi-squared test, RSS test and BF test give the correct answers under independence or dependence respectively. For the six design points under independence, the interquartile ranges are much narrower than their counterparts under dependence. Looking at the modes under dependence, according to Kass and Raftery (1995) there is ‘very strong’ evidence for dependence. We also observe that while there are changes in the magnitudes of the p-values and the log-Bayes factors, the changes in inference over the design points are small whether the modes or the p-values are used. However, there are two exceptions at (ind=1, $\rho = .30$) for $\Delta = 2$ and $\Delta = 4$. Here the

modes are -1.48 and $-.73$, very weak evidence for independence, but there is very strong evidence for independence at $\Delta = 2$, $\rho = .01$ (the mode is -5.86) and $\Delta = 4$, $\rho = .01$ (the mode is -7.50).

However, we observe a few interesting things. First, under independence as ρ increases, both p-values decrease, changing the evidence against independence, but under dependence these p-values increase which changes the degree of evidence against independence (note the minor aberration at (Ind=2, $\Delta = 2$)). However, there is a clear advantage in using the mode because it is the most plausible value, as there is a measure of uncertainty (e.g., the interquartile range) and symmetry between the “association” and “no association” cases. Unlike the behavior of the p-values, the evidence against independence increases as ρ increases (for fixed Ind and δ) for both cases.

In Figure 2.2, we show the distributions of the estimated Bayes factors for the twelve design points. The distributions are essentially unimodal; the locations of the modes tell us about the strength of the evidence against independence. We expect the evidence to be weak under independence but the intra-cluster correlation blurs this vision. As the intraclass correlation increases, we expect more spread, of course, and what we see is that the distributions move over to the right. There is not much change in the distributions with the table density. Also, as we go from independence to dependence, the distributions of the Bayes factor tend to be flatter with more spread.

Finally, we have performed an additional simulation study for a small sample size and a large intraclass correlation. Specifically, we have taken $n = 50$ and $\rho = .50$ with $\ell = 35$ and ind = 1 for independence and ind = 5 for strong dependence. For ind = 5 most of the counts will be on the diagonal of the 3×3 categorical table with the off-diagonal elements tending to be less than 5 and sometimes zero. For ind = 1 some cells will have counts less than 5 because of the strong cluster effect. For ind = 1 the mode and the quartiles of the log-Bayes factor are $-1.38, -1.55, -0.11, 2.00$ and the p-values of the X^2 and RSS tests are, respectively, $.625$ and $.454$. For ind = 5 the mode and the quartiles of the log-Bayes factor are $4.18, 3.02, 6.58, 10.80$ and the p-values of the X^2 and RSS tests are, respectively, 12.73×10^{-5} and 2.06×10^{-5} . Clearly, the RSS test is not able to accommodate the cluster effect with

such a small sample size because the p-value of the RSS test is smaller than that of the X^2 test in both examples.

2.4 Concluding Remarks

We have proposed a method to test for independence in a $r \times c$ contingency table which is obtained from a two-stage cluster sampling design with simple random sampling at both stages. We have used a hierarchical Bayesian model and a sampling-based method to fit it. By making close approximations to several densities we avoid using Markov chain Monte Carlo methods for inference. Specifically, we use random samples from the approximate posterior density and subsample them using the SIR algorithm. Although ours is a sampling based method it is at least as fast as the Rao-Scott methods. We use the Bayes factor to make inference about independence. Relative to standard methods our approach provides additional insight by displaying the distribution of the Bayes factor rather than simply relying on a single summary measure.

The Rao-Scott methods were developed to correct for design effects such as cluster effects, i.e., by correcting the standard X^2 and G^2 statistics. They are “large sample” methods and work well when there are large cell counts. However, they are less successful when there are small cell counts. An extreme case is a table with zero counts, in which case the X^2 and G^2 tests are not applicable. Consequently, the Rao-Scott methods do not apply either (since they are adjustments of the X^2 and G^2 tests for design effects, not sparse tables). Our procedure will get around this problem when there are a few cells having zero counts. However, by doing a more sophisticated analysis, we have validated RSS for two-stage cluster sampling with many examples, but as we discussed, there are some examples when this is not quite true.

Finally, we note that in small complex surveys, most cluster tables will have many zero cells (e.g., contingency tables with categorical variables having many levels). As noted above the problem of sparse total tables cannot be accommodated within the Rao-Scott framework. However, it may be possible to do so within our framework. For example, a likelihood ratio test of independence in a single contingency table with many sampling zeros is given by

Nandram, Bhatta and Bhadra (2012) assuming simple random sampling. It will be useful to extend this work to complex surveys.

Table 2.1: Features of the total table for each of the eight examples

	n	ℓ	ρ	def	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)	(3,1)	(3,2)	(3,3)
E1	469	37	.56	11.8	44	57	5	83	71	5	63	136	5
E2	663	24	.33	6.85	49	74	1	107	151	13	93	164	11
E3	438	23	.34	7.39	44	47	8	54	44	3	56	167	15
E4	857	51	.33	6.62	25	17	0	157	134	13	205	294	12
E5	469	24	.54	11.46	63	38	5	105	47	7	70	124	10
E6	663	37	.31	6.45	61	56	7	117	141	13	117	145	6
E7	438	23	.35	7.67	53	44	2	67	30	4	95	133	10
E8	857	51	.33	6.63	34	7	1	181	112	11	226	272	13

NOTE: These are all 3×3 contingency tables; n is the number of observations; ℓ is the number of schools; ρ is the intracluster correlation and Def stands for design effect. E4 has a zero cell and E2, E3, E7, E8 have some cell counts near zero.

Table 2.2: Bayesian Design effects for each cell by example

Cell	E1	E2	E3	E4	E5	E6	E7	E8
(1,1)	7.42	5.17	4.75	5.32	6.88	4.93	4.72	5.12
(1,2)	6.99	5.09	4.92	5.69	6.66	4.59	5.25	7.45
(1,3)	12.10	17.05	7.35	25.65	12.69	7.51	12.91	17.59
(2,1)	7.94	5.31	4.81	5.22	7.65	5.07	4.83	5.31
(2,2)	6.88	5.18	5.47	5.04	6.37	5.06	5.35	5.13
(2,3)	14.04	5.85	9.99	6.14	11.40	6.13	9.43	6.59
(3,1)	6.23	5.20	5.23	5.37	6.45	5.02	5.92	5.39
(3,2)	6.54	5.35	4.85	5.14	6.59	5.30	5.27	5.22
(3,3)	12.97	6.56	5.80	6.60	10.37	7.40	6.70	6.09
ESS	67	126	87	164	68	130	82	161

NOTE: The cells are (j, k) , $j, k = 1, 2, 3$. ESS stands for the effective sample size and it is the sum of the cell counts divided by the design effects, taken for the total table.

Table 2.3: Comparison of the log-Bayes factor with the p-values by example

	p-values			log-Bayes factor							
	χ^2	RSF	RSS	Min	Q_1	Q_2	Q_3	Max	Mode	P	N
E1	.001	.17	.14	-7.1	3.6	12.5	23.3	105	5.7	.81	9
E2	.247	.58	.66	-7.9	-2.8	1.1	7.2	60	-1.6	.89	9
E3	.000	.04	.02	-7.4	7.9	17.0	28.0	94	10.8	.69	7
E4	.001	.04	.02	-8.6	1.0	8.0	16.9	84	4.5	.76	9
E5	.000	.02	.01	-6.4	9.3	20.1	33.7	109	14.6	.66	7
E6	.240	.60	.69	-7.9	-1.3	3.4	10.2	55	-0.7	.95	9
E7	.000	.03	.01	-7.2	2.5	9.6	18.5	77	6.05	.71	9
E8	.000	.01	.00	-8.1	7.2	16.2	26.6	108	10.6	.66	7

NOTE: RSF and RSS denote, respectively, the first and second order Rao-Scott approximations; a bootstrap method is used to estimate the covariance matrix in the Rao-Scott approximations.

Table 2.4: Sensitivity analysis of the log-Bayes factor with respect to $\tau_s, s = 1, \dots, 9$, by region (reg) and example

reg	η	MTS vs. COM			STS vs. COM		
		.5	1	2	.5	1	2
1	Mode	9.3	6.3	4.9	24.5	12.6	9.4
	Median	13.8	12.4	10.3	25.6	21.1	15.6
	IQR	(4.8,25.4)	(4.1,23.5)	(2.9,19.0)	(11.9,37.8)	(9.9, 33.7)	(6.9, 26.9)
	P	.84	.82	.79	.73	.66	.56
2	Mode	-1.5	-1.7	-2.6	-0.9	-0.5	-0.5
	Median	2.1	1.1	1.0	3.6	3.5	2.8
	IQR	(-2.3,7.7)	(-2.5,7.3)	(-3.1,6.9)	(-1.5,10.8)	(-1.5,10.7)	(-1.7,9.1)
	P	.90	.90	.88	.95	.94	.94
3	Mode	14.0	11.1	10.1	5.8	4.9	2.5
	Median	17.5	16.5	13.9	10.5	9.6	6.8
	IQR	(8.3,28.6)	(7.5,27.0)	(6.1,23.1)	(3.6,21.2)	(3.0,18.7)	(1.0,14.8)
	P	.70	.68	.62	.74	.72	.62
4	Mode	3.7	4.2	1.8	12.6	12.1	9.2
	Median	8.8	8.2	7.0	17.0	15.7	14.0
	IQR	(1.8,19.0)	(1.6,16.9)	(0.4,15.9)	(8.3,28.4)	(6.6,28.0)	(5.1,24.1)
	P	.78	.78	.73	.70	.66	.62

NOTE: Each region has two examples (e.g., region 1 corresponds to E1, left, and E5, right).

We have used $\tau_s = \eta \hat{\tau}_s, s = 1, \dots, 9$, where the τ_s are maximum likelihood estimates and $\eta = .5, 1, 2$.

Table 2.5: Simulation: Comparison of p-values and log-Bayes factor

Ind	Δ	ρ	Def	p-values		log-Bayes factor			
				χ^2	RSS	mode	Q_1	Q_2	Q_3
1	2	.01	1.00	.964	.993	-5.86	-6.14	-4.50	-2.31
1	2	.10	1.50	.760	.899	-4.66	-5.35	-3.36	-0.51
1	2	.30	3.74	.350	.816	-1.48	-2.78	0.85	6.06
1	4	.01	1.00	.991	.999	-7.50	-7.61	-6.08	-3.95
1	4	.10	2.32	.700	.883	-5.62	-5.97	-3.25	0.38
1	4	.30	6.71	.280	.884	-0.73	-1.76	3.46	11.39
2	2	.01	1.15	.005	.001	6.16	2.47	8.44	15.58
2	2	.10	2.70	.006	.010	7.60	4.73	12.09	23.47
2	2	.30	6.12	.001	.044	11.19	8.23	19.87	35.79
2	4	.01	1.29	.000	.000	15.40	11.83	20.33	31.35
2	4	.10	4.38	.000	.002	25.39	15.33	30.57	47.62
2	4	.30	11.26	.001	.033	29.50	19.14	39.87	66.92

NOTE: Ind is the degree of dependence, Δ is table density, ρ is the intracluster correlation, Def is the design effect from Briers method, RSS is the second order Rao-Scott correction.

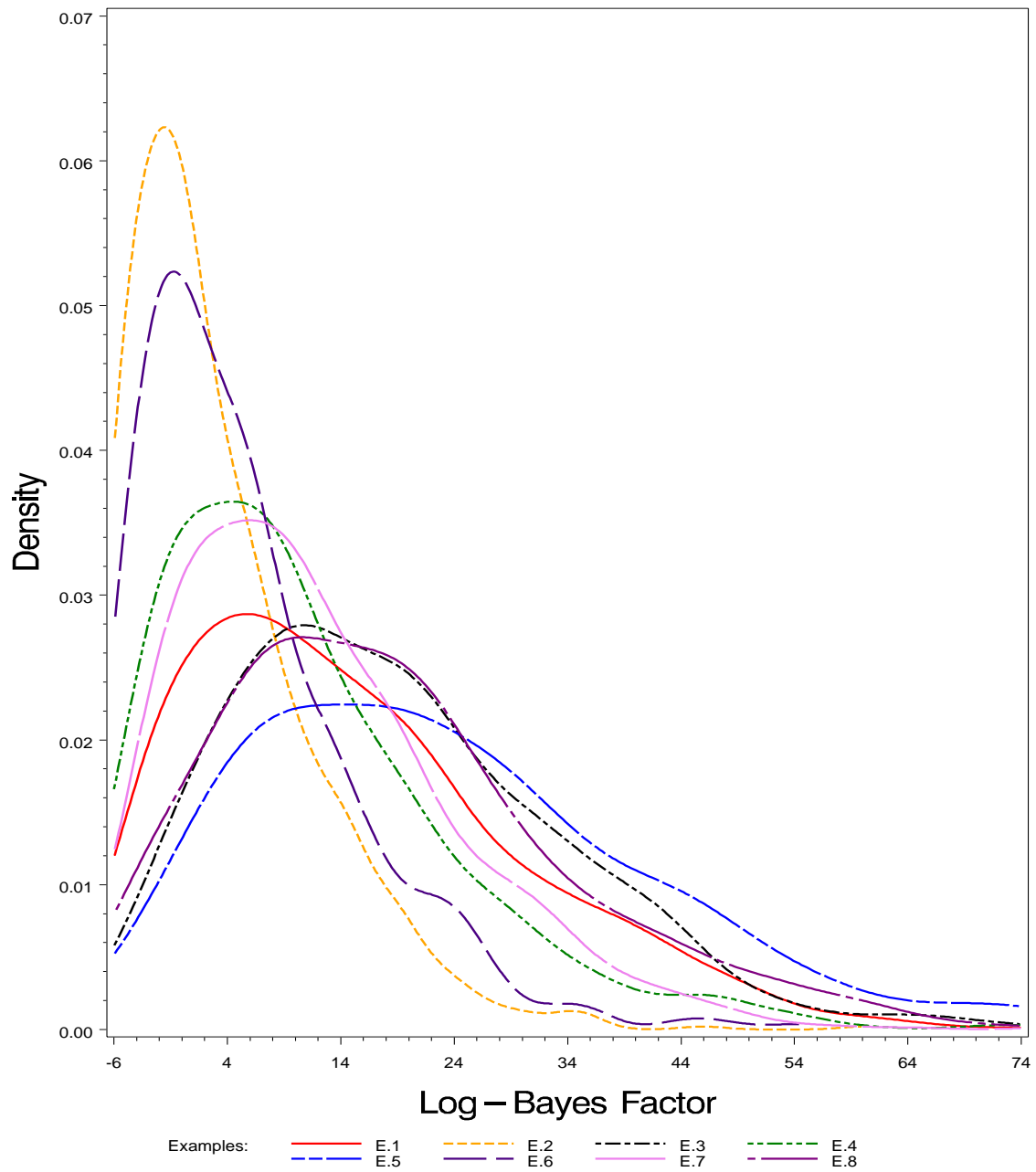


Figure 2.1: Plots of the empirical densities of the log-Bayes factors for the eight strata in the third grade example

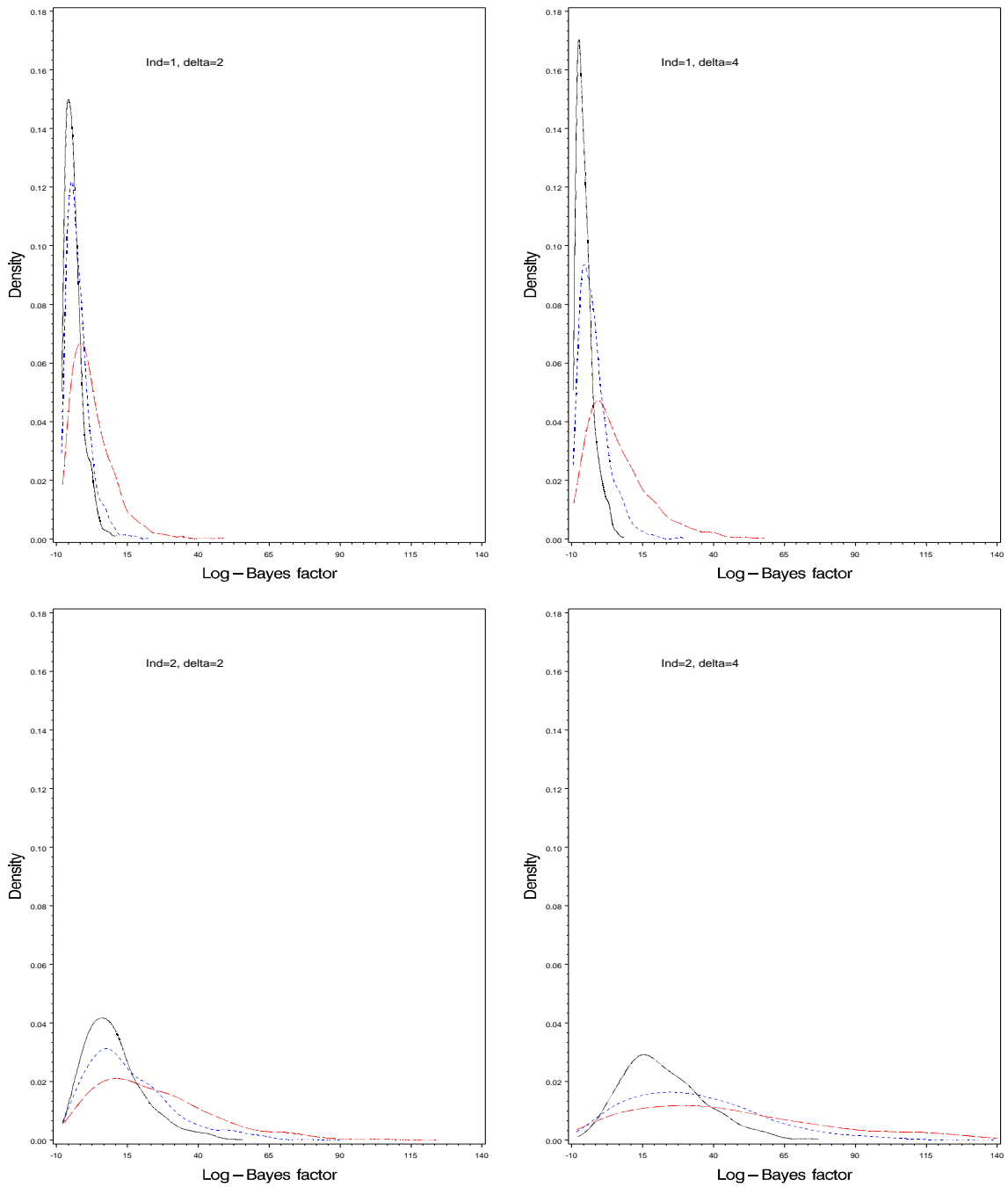


Figure 2.2: Simulation: Plots of the empirical densities of the log-Bayes factors at twelve design points. The symbols are correlation (solid: $\rho = .01$, dotted: $\rho = .10$, long dashed: $\rho = .30$), association (independence: ind=1 and dependence: ind=2) and table density (Δ)

Chapter 3

A Test of Independence With Covariates

In Chapter 3, we discuss the test of independence in a two-way contingency table when there are covariates at both unit and cluster levels. These covariates are likely to be associated with the two cross-classified categorical variables and can have influence over their association. If we simply ignore the effect of covariates and perform the test of independence, the test can be misleading. Geenens and Simar (2010, 2011) developed nonparametric and semiparametric methods for conditional independence in two-way contingency tables for simple random sampling. However, we did not find any literature for the test of independence on complex survey data when there are covariates.

To perform the Bayesian test of independence, we use the idea of surrogate sampling similar to the one applied in Chapter 2. However, in this case, we use a two-step procedure to compute the Bayes factor by using a surrogate sample. First, we fit a multinomial logistic regression model with random effects to the observed cluster data with covariates. The random effect in the model accommodates the cluster effect in the data and the covariates incorporated in the model explain the fixed effect. After fitting the random effects model, we predict the samples given the set of covariates. The sample obtained as such would be the cluster sample without covariates because the sample has already been adjusted with the covariates. Then, in the next step we use a hierarchical Bayesian model to convert the cluster sample without covariates into an equivalent simple random sample.

Here, we have developed a new methodology instead of using the same one as Chapter 2. In this new situation, the method we adopted in Chapter 2 is computationally expensive.

First, because we use SIR algorithm in Chapter 2 which requires a large number of samples to subsample. Second, we have many constraints in the model from Chapter 2 which makes computation complicated. However, we have fewer constraints if we use logistic regression because many constraints are automatically incorporated there through its structure.

In Chapter 3, we will discuss seven more sections. In Section 3.1, we describe a random effects multinomial logistic regression model. In Section 3.2, we describe the cluster model without covariates. In Section 3.3, we show how to compute the Bayes factor from the surrogate samples. In Section 3.4, we present an example (TIMSS 2007 data) which we use to illustrate our methodology. We also fit our cluster model without covariates in order to make a comparison between the test with covariates and the test without covariates. In addition, we will also study the effect of covariates on the test. In Section 3.5, we perform a simulation study to validate the findings we obtain for the real data. In Section 3.6, we study the empirical power function of our Bayes factor test statistic. Finally, Section 3.7 has concluding remarks.

3.1 A Random Effect Multinomial Logistic Regression Model

Consider an $r \times c$ categorical table of a sample of n_i individuals for the i^{th} , $i = 1, \dots, \ell$, cluster. Also consider a set of individual (or unit) level covariates X and the set of cluster level covariates Z . We string out the observations in a table to an array of $S = rc$ cells. Let $I_{ijs} = 1$ if the j^{th} ssu (or individual) falls in the s^{th} cell within i^{th} cluster and $I_{ijs} = 0$ otherwise. Then,

$$I_{ij} \stackrel{ind}{\sim} \text{Multinomial}(1, a_{ij}), \quad i = 1, \dots, \ell, \quad j = 1, \dots, n_i, \quad (3.1)$$

where

$$a_{ijs} = \begin{cases} \frac{e^{\underline{\beta}_s' \underline{x}_{ij} + \underline{\gamma}' \underline{z}_{i+\delta_i}}}{1 + \sum_{s=1}^{S-1} e^{\underline{\beta}_s' \underline{x}_{ij} + \underline{\gamma}' \underline{z}_{i+\delta_i}}}, & s = 1, \dots, S-1 \\ \frac{1}{1 + \sum_{s=1}^{S-1} e^{\underline{\beta}_s' \underline{x}_{ij} + \underline{\gamma}' \underline{z}_{i+\delta_i}}}, & s = S, \end{cases}$$

and $\delta_i \stackrel{iid}{\sim} N(0, \sigma^2)$, $i = 1, \dots, \ell$, are the random effects. Here, $\underline{\beta}_s$ is the $p \times 1$ vector of regression parameters, \underline{x}_{ij} the $p \times 1$ vector of unit level covariates and similarly $\underline{\gamma}$ and \underline{z}_i are

the $q \times 1$ vectors of regression parameter and cluster level covariates respectively. Then, the likelihood of an individual falling in cells $1, \dots, S$ is

$$P(\underline{I}_{ij} | \underline{\beta}, \underline{\gamma}, \sigma^2) = \frac{\prod_{s=1}^{S-1} e^{(\underline{\beta}_s' \underline{x}_{ij} + \underline{\gamma}' z_i + \delta_i) I_{ijs}}}{1 + \sum_{s=1}^{S-1} e^{(\underline{\beta}_s' \underline{x}_{ij} + \underline{\gamma}' z_i + \delta_i) I_{ijs}}}.$$

For computational convenience we take $\nu_i = \underline{\gamma}' z_i + \delta_i$ so that $\nu_i | \underline{\gamma}, \sigma^2 \stackrel{ind}{\sim} N(\underline{\gamma}' z_i, \sigma^2)$.

Then, the likelihood function becomes

$$P(\underline{I} | \underline{\beta}, \underline{\gamma}, \sigma^2) \propto \prod_{i=1}^{\ell} \prod_{j=1}^{n_i} \left\{ \frac{\prod_{s=1}^{S-1} e^{(\underline{\beta}_s' \underline{x}_{ij} + \nu_i) I_{ijs}}}{1 + \sum_{s=1}^{S-1} e^{\underline{\beta}_s' \underline{x}_{ij} + \nu_i}} \right\}, \quad (3.2)$$

where $\underline{I} = \{I_{ij}, i = 1, \dots, \ell, j = 1, \dots, n_i\}$.

A priori, we take

$$\pi(\underline{\beta}) \propto 1, \quad \underline{\gamma} | \sigma^2 \sim N_q(\underline{\gamma}_0, \sigma^2 \Delta_0) \quad \text{and} \quad \sigma^2 \sim \text{IGamma}(a/2, b/2), \quad (3.3)$$

where we choose $a = b = .001$. we will also study the sensitivity of these specification. We take $\underline{\gamma}_0 = \hat{\underline{\gamma}}$ and $\Delta_0 = \kappa \hat{\Delta} = \kappa (z' z)^{-1}$, where $\hat{\underline{\gamma}}$ and $\hat{\Delta}$ are the maximum likelihood estimators of $\underline{\gamma}$ and Δ , the covariance matrix of $\underline{\gamma}$, and z is a $\ell \times q$ design matrix of the cluster covariates. Note here that we can use any value for $\underline{\gamma}_0$ not only $\hat{\underline{\gamma}}$. We take κ large enough ($\kappa = 100$) so that the prior is proper diffuse.

3.1.1 The Joint Posterior Density

Combining the likelihood in (3.2) and the priors in (3.3) via Bayes' theorem, given the sample data \underline{I} , we get the joint posterior density of $\underline{\beta}, \underline{\nu}, \underline{\gamma}, \sigma^2$

$$\begin{aligned} \pi(\underline{\beta}, \underline{\nu}, \underline{\gamma}, \sigma^2 | \underline{I}) &\propto \prod_{i=1}^{\ell} \prod_{j=1}^{n_i} \left\{ \frac{\prod_{s=1}^{S-1} e^{(\underline{\beta}_s' \underline{x}_{ij} + \nu_i) I_{ijs}}}{1 + \sum_{s=1}^{S-1} e^{\underline{\beta}_s' \underline{x}_{ij} + \nu_i}} \right\} \times \prod_{i=1}^{\ell} \left\{ (1/\sigma^2)^{1/2} e^{-\frac{1}{2\sigma^2} (\nu_i - \underline{\gamma}' z_i)^2} \right\} \\ &\quad \times \frac{1}{(\sigma^2)^{q/2} |\Delta_0|^{1/2}} e^{-\frac{1}{2\sigma^2} (\underline{\gamma} - \underline{\gamma}_0)' \Delta_0^{-1} (\underline{\gamma} - \underline{\gamma}_0)} \times (1/\sigma^2)^{a/2-1} e^{-b/2\sigma^2} \\ &= \prod_{i=1}^{\ell} \prod_{j=1}^{n_i} \left\{ \frac{\prod_{s=1}^{S-1} \prod_{k=0}^p (e^{\beta_{sk}})^{x_{ijk} I_{ijs}} (e^{\nu_i})^{\sum_{s=1}^{S-1} I_{ijs}}}{1 + e^{\nu_i} \sum_{s=1}^{S-1} \prod_{k=0}^p (e^{\beta_{sk}})^{x_{ijk}}} \right\} \times \prod_{i=1}^{\ell} \left\{ (1/\sigma^2)^{1/2} e^{-\frac{1}{2\sigma^2} (\nu_i - \underline{\gamma}' z_i)^2} \right\} \\ &\quad \times \frac{1}{(\sigma^2)^{q/2} |\Delta_0|^{1/2}} e^{-\frac{1}{2\sigma^2} (\underline{\gamma} - \underline{\gamma}_0)' \Delta_0^{-1} (\underline{\gamma} - \underline{\gamma}_0)} \times (1/\sigma^2)^{a/2-1} e^{-b/2\sigma^2}. \end{aligned} \quad (3.4)$$

We transform each β_{sk} from $(-\infty, \infty)$ to $(0, 1)$. The transformation is useful because it reduces the complexity of the computation and the fact that $0 < \phi_{sk} < 1$, it helps to do the proof of propriety.

$$e^{\beta_{sk}} = \phi_{sk}/(1 - \phi_{sk}), \quad s = 1, \dots, S-1, \quad k = 0, \dots, p, \quad (3.5)$$

ϕ_{sk} are now in $(0, 1)$ and the Jacobian of the transformation is

$|J| = \prod_{s=1}^{S-1} \prod_{k=0}^p 1/\{\phi_{sk}(1 - \phi_{sk})\}$. Then,

$$\begin{aligned} \pi(\underline{\phi}, \underline{\nu}, \underline{\gamma}, \sigma^2 \mid \underline{I}) &\propto \prod_{i=1}^{\ell} \left\{ \frac{\prod_{s=1}^{S-1} \prod_{k=0}^p \left(\frac{\phi_{sk}}{1-\phi_{sk}}\right)^{\sum_{j=1}^{n_i} x_{ijk} I_{ijs}} (e^{\nu_i})^{\sum_{j=1}^{n_i} \sum_{s=1}^{S-1} I_{ijs}}}{\prod_{j=1}^{n_i} [1 + e^{\nu_i} \sum_{s=1}^{S-1} \prod_{k=0}^p \left(\frac{\phi_{sk}}{1-\phi_{sk}}\right)^{x_{ijk}}]} \right\} \\ &\times \prod_{i=1}^{\ell} \left\{ (1/\sigma^2)^{1/2} e^{-\frac{1}{2\sigma^2}(\nu_i - \underline{\gamma}' \underline{z}_i)^2} \right\} \times \frac{1}{(\sigma^2)^{q/2} |\Delta_0|^{1/2}} e^{-\frac{1}{2\sigma^2}(\underline{\gamma} - \underline{\gamma}_0)' \Delta_0^{-1} (\underline{\gamma} - \underline{\gamma}_0)} \\ &\times (1/\sigma^2)^{a/2-1} e^{-b/2\sigma^2} \times \left\{ \prod_{s=1}^{S-1} \prod_{k=0}^p \frac{1}{\phi_{sk}(1 - \phi_{sk})} \right\}. \end{aligned} \quad (3.6)$$

Letting $a_i = \sum_{j=1}^{n_i} \sum_{s=1}^{S-1} I_{ijs}$, the posterior density in (3.6) becomes

$$\begin{aligned} \pi(\underline{\phi}, \underline{\nu}, \underline{\gamma}, \sigma^2 \mid \underline{I}) &\propto \prod_{i=1}^{\ell} \left\{ \frac{\prod_{s=1}^{S-1} \prod_{k=0}^p \left(\frac{\phi_{sk}}{1-\phi_{sk}}\right)^{\sum_{j=1}^{n_i} x_{ijk} I_{ijs}} e^{\nu_i a_i}}{\prod_{j=1}^{n_i} [1 + e^{\nu_i} \sum_{s=1}^{S-1} \prod_{k=0}^p \left(\frac{\phi_{sk}}{1-\phi_{sk}}\right)^{x_{ijk}}]} \right\} \times (1/\sigma^2)^{(\ell+a+q)/2-1} \\ &\times e^{-\frac{1}{2\sigma^2} [b + \sum_{i=1}^{\ell} (\nu_i - \underline{\gamma}' \underline{z}_i)^2 + (\underline{\gamma} - \underline{\gamma}_0)' \Delta_0^{-1} (\underline{\gamma} - \underline{\gamma}_0)]} \\ &\times \left\{ \prod_{s=1}^{S-1} \prod_{k=0}^p \frac{1}{\phi_{sk}(1 - \phi_{sk})} \right\}. \end{aligned} \quad (3.7)$$

We would like to simplify the term $\sum_{i=1}^{\ell} (\nu_i - \underline{\gamma}' \underline{z}_i)^2 + (\underline{\gamma} - \underline{\gamma}_0)' \Delta_0^{-1} (\underline{\gamma} - \underline{\gamma}_0)$ in (3.7) further. Maximizing the likelihood function of $\nu_i \stackrel{ind}{\sim} N(\underline{\gamma}' \underline{z}_i, \sigma^2)$ with respect to γ , we get $\hat{\gamma} = (\sum_{i=1}^{\ell} \underline{z}_i \underline{z}_i')^{-1} (\sum_{i=1}^{\ell} \nu_i \underline{z}_i) = (\underline{z}' \underline{z})^{-1} (\underline{z}' \underline{\nu})$, the maximum likelihood estimator of γ . We can write

$$\begin{aligned} \sum_{i=1}^{\ell} (\nu_i - \underline{\gamma}' \underline{z}_i)^2 + (\underline{\gamma} - \underline{\gamma}_0)' \Delta_0^{-1} (\underline{\gamma} - \underline{\gamma}_0) &= \sum_{i=1}^{\ell} (\nu_i - \hat{\gamma}' \underline{z}_i + \hat{\gamma}' \underline{z}_i - \underline{\gamma}' \underline{z}_i)^2 + (\underline{\gamma} - \underline{\gamma}_0)' \Delta_0^{-1} (\underline{\gamma} - \underline{\gamma}_0) \\ &= \sum_{i=1}^{\ell} (\nu_i - \hat{\gamma}' \underline{z}_i)^2 + (\hat{\gamma} - \underline{\gamma})' \hat{\Delta}^{-1} (\hat{\gamma} - \underline{\gamma}) + (\underline{\gamma} - \underline{\gamma}_0)' \Delta_0^{-1} (\underline{\gamma} - \underline{\gamma}_0), \end{aligned} \quad (3.8)$$

where $\hat{\Delta} = (z'z)^{-1}$. We have simplified (3.8) further in Appendix D to get

$$\begin{aligned} & \sum_{i=1}^{\ell} (\nu_i - \gamma' z_i)^2 + (\gamma - \gamma_0)' \Delta_0^{-1} (\gamma - \gamma_0) \\ &= \gamma_0' \frac{(z'z)}{\kappa+1} \gamma_0 + \nu' [I - \frac{\kappa}{\kappa+1} z(z'z)^{-1} z'] \nu - \frac{2}{\kappa+1} \gamma_0' z' \nu \\ & \quad + (\gamma - \frac{\kappa \hat{\gamma} + \gamma_0}{\kappa+1})' (\frac{\kappa+1}{\kappa}) \hat{\Delta}^{-1} (\gamma - \frac{\kappa \hat{\gamma} + \gamma_0}{\kappa+1}). \end{aligned}$$

Thus, we can write the distribution (3.7) as

$$\begin{aligned} \pi(\underline{\phi}, \underline{\nu}, \gamma, \sigma^2 | \underline{I}) &\propto \prod_{i=1}^{\ell} \left\{ \frac{\prod_{s=1}^{S-1} \prod_{k=0}^p \left(\frac{\phi_{sk}}{1-\phi_{sk}} \right)^{\sum_{j=1}^{n_i} x_{ijk} I_{ijs}} e^{\nu_i a_i}}{\prod_{j=1}^{n_i} \left[1 + e^{\nu_i} \sum_{s=1}^{S-1} \prod_{k=0}^p \left(\frac{\phi_{sk}}{1-\phi_{sk}} \right)^{x_{ijk}} \right]} \right\} \times (1/\sigma^2)^{(\ell+a+q)/2-1} \\ &\times e^{-\frac{1}{2\sigma^2} \left\{ b + \gamma_0' \frac{(z'z)}{\kappa+1} \gamma_0 + \nu' [I - \frac{\kappa}{\kappa+1} z(z'z)^{-1} z'] \nu \right.} \\ &\quad \left. - \frac{2}{\kappa+1} \gamma_0' z' \nu + (\gamma - \frac{\kappa \hat{\gamma} + \gamma_0}{\kappa+1})' (\frac{\kappa+1}{\kappa}) \hat{\Delta}^{-1} (\gamma - \frac{\kappa \hat{\gamma} + \gamma_0}{\kappa+1}) \right\}} \\ &\times \prod_{s=1}^{S-1} \prod_{k=0}^p \frac{1}{\phi_{sk} (1 - \phi_{sk})}. \end{aligned} \quad (3.9)$$

Letting

$$g(\phi) = \prod_{s=1}^{S-1} \prod_{k=0}^p \left\{ \left(\frac{\phi_{sk}}{1 - \phi_{sk}} \right)^{\sum_{i=1}^{\ell} \sum_{j=1}^{n_i} x_{ijk} I_{ijs}} \times \frac{1}{\phi_{sk} (1 - \phi_{sk})} \right\}, \quad (3.10)$$

and

$$h(\nu_i) = \frac{e^{\nu_i a_i}}{\prod_{j=1}^{n_i} \left\{ 1 + e^{\nu_i} \sum_{s=1}^{S-1} \prod_{k=0}^p \left(\frac{\phi_{sk}}{1-\phi_{sk}} \right)^{x_{ijk}} \right\}}, \quad (3.11)$$

we can write the distribution in (3.9) as

$$\begin{aligned} \pi(\underline{\phi}, \underline{\nu}, \gamma, \sigma^2 | \underline{I}) &\propto g(\underline{\phi}) \left\{ \prod_{i=1}^{\ell} h(\nu_i) \right\} (1/\sigma^2)^{(\ell+a+q)/2-1} \\ &\times e^{-\frac{1}{2\sigma^2} \left\{ b + \gamma_0' \frac{(z'z)}{\kappa+1} \gamma_0 + \nu' [I - \frac{\kappa}{\kappa+1} z(z'z)^{-1} z'] \nu \right.} \\ &\quad \left. - \frac{2}{\kappa+1} \gamma_0' z' \nu + (\gamma - \frac{\kappa \hat{\gamma} + \gamma_0}{\kappa+1})' (\frac{\kappa+1}{\kappa}) \hat{\Delta}^{-1} (\gamma - \frac{\kappa \hat{\gamma} + \gamma_0}{\kappa+1}) \right\}. \end{aligned} \quad (3.12)$$

We next develop some theoretical properties about the joint posterior density in (3.12).

Lemma 3.1.1. *Each univariate function $h(\nu_i)$ in (3.11) is log-concave.*

Proof: Let

$$c_{ij} = \sum_{s=1}^{S-1} \prod_{k=0}^p \left(\frac{\phi_{sk}}{1 - \phi_{sk}} \right)^{x_{ijk}}. \quad (3.13)$$

Then, we have from (3.11)

$$h(\nu_i) = \frac{e^{a_i \nu_i}}{\prod_{j=1}^{n_i} \{1 + e^{\nu_i c_{ij}}\}}. \quad (3.14)$$

Taking the logarithm of both sides

$$\Delta = \log[h(\nu_i)] = a_i \nu_i - \sum_{j=1}^{n_i} \log(1 + e^{\nu_i c_{ij}}).$$

Then, the first and second order derivatives are

$$\begin{aligned} \Delta' &= a_i - \sum_{j=1}^{n_i} \frac{e^{\nu_i c_{ij}}}{1 + e^{\nu_i c_{ij}}} \\ \Delta'' &= - \sum_{j=1}^{n_i} \frac{e^{\nu_i c_{ij}}}{(1 + e^{\nu_i c_{ij}})^2}. \end{aligned}$$

Because $e^{\nu_i} > 0$ and $c_{ij} > 0$, $\Delta'' < 0$. Therefore, $h(\nu_i)$ is log-concave. This allows us to use adaptive rejection sampling (ARS) to draw ν_i from its conditional posterior density.

Now, we are going to show that joint posterior density in (3.12) is proper under very mild conditions. We consider the case when $\kappa \rightarrow \infty$ under which the prior of γ becomes improper, the worst case. Then, the joint posterior density is

$$\begin{aligned} \pi(\underline{\phi}, \underline{\nu}, \underline{\gamma}, \sigma^2 \mid \underline{I}) &\propto g(\underline{\phi}) \left\{ \prod_{i=1}^{\ell} h(\nu_i) \right\} (1/\sigma^2)^{(\ell+a+q)/2-1} \\ &\times e^{-\frac{1}{2\sigma^2} \{b + \underline{\nu}' [I - z(z'z)^{-1}z'] \underline{\nu} + (\underline{\gamma} - \hat{\underline{\gamma}})' \hat{\Delta}^{-1} (\underline{\gamma} - \hat{\underline{\gamma}})\}}. \end{aligned} \quad (3.15)$$

Theorem 3.1.1. *For $0 < \phi_{sk} < 1$, the joint posterior density in (3.15) is proper.*

Proof: Integrating out γ (which has multivariate normal) from (3.15) we get

$$\begin{aligned} \pi(\underline{\phi}, \underline{\nu}, \sigma^2 \mid \underline{I}) &\propto g(\underline{\phi}) \left\{ \prod_{i=1}^{\ell} h(\nu_i) \right\} (1/\sigma^2)^{(\ell+a)/2-1} \\ &\times e^{-\frac{1}{2\sigma^2} [b + \underline{\nu}' [I - z(z'z)^{-1}z'] \underline{\nu}]}. \end{aligned} \quad (3.16)$$

Let $P = I - z(z'z)^{-1}z'$ be a projection matrix. It is not invertible because $\text{rank}(P) = \ell - q$. By Seber (1984), for any projection matrix of rank r we can write $P = \sum_{i=1}^r t_i t_i'$, where t_1, \dots, t_r form an orthonormal set. Thus, we can write $\underline{\nu}' [I - z(z'z)^{-1}z'] \underline{\nu} = \underline{\nu}' P \underline{\nu} =$

$\sum_{i=1}^{\ell-q} (t_i' \underline{\nu})' t_i' \underline{\nu}$. Now, let us make a one to one transformation as

$$\begin{pmatrix} y_1 \\ \vdots \\ y_{\ell-q} \\ y_{\ell-q+1} \\ \vdots \\ y_\ell \end{pmatrix} = \begin{pmatrix} t'_1 \\ \vdots \\ t'_{\ell-q} \\ 0 \cdots 0 & 1 \cdots 0 \\ \vdots \cdots \vdots & \vdots \ddots \vdots \\ 0 \cdots 0 & 0 \cdots 1 \end{pmatrix} \begin{pmatrix} \nu_1 \\ \vdots \\ \nu_{\ell-q} \\ \nu_{\ell-q+1} \\ \vdots \\ \nu_\ell \end{pmatrix}$$

so that $\underline{y} = A\underline{\nu}$. Here, we keep $\nu_{\ell-q+1}, \dots, \nu_\ell$ as untransformed. Since A is invertible, we have $\underline{\nu} = A^{-1}\underline{y}$. Let us denote $\underline{y}^{(1)'} = [y_1, \dots, y_{\ell-q}]$, $\underline{y}^{(2)'} = [y_{\ell-q+1}, \dots, y_\ell]$, $\underline{\nu}^{(1)'} = [\nu_1, \dots, \nu_{\ell-q}]$, and $\underline{\nu}^{(2)'} = [\nu_{\ell-q+1}, \dots, \nu_\ell]$, then, $\underline{\nu}'[I - z(z'z)^{-1}z']\underline{\nu} = \sum_{j=1}^{\ell-q} \nu_j^2 = \underline{y}^{(1)'}\underline{y}^{(1)}$. Note here that $\underline{y}^{(2)}$ and $\underline{\nu}^{(2)}$ are the same vector. The transformation is one to one, so the Jacobian matrix is not a function of any of the random variables and contains only constant elements. With this transformation, the joint posterior in (3.16) becomes

$$\begin{aligned} \pi(\underline{\phi}, \underline{y}^{(1)}, \underline{\nu}^{(2)}, \sigma^2 \mid \underline{I}) &\propto g(\underline{\phi}) \left\{ \prod_{i=1}^{\ell-q} h(y_i^*) \right\} \left\{ \prod_{i=\ell-q+1}^{\ell} h(\nu_i) \right\} (1/\sigma^2)^{(\ell+a)/2-1} \\ &\times e^{-\frac{1}{2\sigma^2} [b + \underline{y}^{(1)'}\underline{y}^{(1)}]}, \end{aligned}$$

where $y_i^* = (A^{-1}\underline{y})_i$. Here, $h(y_i^*)$ is bounded by its maximum $B_i(\underline{\phi})$ due to the log-concavity of $h(\nu_i)$ as proved in Lemma 3.1.1. This means that $\prod_{i=1}^{\ell-q} h(y_i^*)$ is bounded by $B(\underline{\phi}) = \max\{B_i(\underline{\phi}), i = 1, \dots, \ell - q\}$. Therefore,

$$\begin{aligned} \pi(\underline{\phi}, \underline{y}^{(1)}, \underline{\nu}^{(2)}, \sigma^2 \mid \underline{I}) &\propto g(\underline{\phi}) \left\{ \prod_{i=1}^{\ell-q} h(y_i^*) \right\} \left\{ \prod_{i=\ell-q+1}^{\ell} h(\nu_i) \right\} (1/\sigma^2)^{(\ell+a)/2-1} \\ &\times e^{-\frac{1}{2\sigma^2} [b + \underline{y}^{(1)'}\underline{y}^{(1)}]} \\ &< g(\underline{\phi}) B(\underline{\phi}) \left\{ \prod_{i=\ell-q+1}^{\ell} h(\nu_i) \right\} (1/\sigma^2)^{(\ell+a)/2-1} \\ &\times e^{-\frac{1}{2\sigma^2} [b + \underline{y}^{(1)'}\underline{y}^{(1)}]}. \end{aligned}$$

Next we show that $\int_{\underline{\phi}} \int_{\underline{y}^{(1)}} \int_{\sigma^2} \int_{\underline{\nu}^{(2)}} \pi(\underline{\phi}, \underline{y}^{(1)}, \sigma^2, \underline{\nu}^{(2)} \mid \underline{I}) d\underline{\nu}^{(2)} d\sigma^2 d\underline{y}^{(1)} d\underline{\phi}$ is finite.

Let us denote

$$\begin{aligned}
& \int_{\underline{\phi}} \int_{\underline{y}^{(1)}} \int_{\sigma^2} \int_{\underline{\nu}^{(2)}} \pi(\underline{\phi}, \underline{y}^{(1)}, \sigma^2, \underline{\nu}^{(2)} \mid \underline{I}) d\underline{\nu}^{(2)} d\sigma^2 d\underline{y}^{(1)} d\underline{\phi} \propto F \\
& = \int_{\underline{\phi}} \int_{\underline{y}^{(1)}} \int_{\sigma^2} \int_{\underline{\nu}^{(2)}} g(\underline{\phi}) B(\underline{\phi}) \left\{ \prod_{i=\ell-q+1}^{\ell} h(\nu_i) \right\} (1/\sigma^2)^{(\ell+a)/2-1} \\
& \times e^{-\frac{1}{2\sigma^2} [b + \underline{y}^{(1)'} \underline{y}^{(1)}]} d\underline{\nu}^{(2)} d\sigma^2 d\underline{y}^{(1)} d\underline{\phi}.
\end{aligned}$$

Note here that we do not want to bound $\prod_{i=\ell-q+1}^{\ell} h(\nu_i)$ otherwise the integration of the function with respect to $\underline{\nu}^{(2)}$ would be infinite. We want to show that F is finite. Here,

$$\begin{aligned}
F & = \int_{\underline{\phi}} \int_{\underline{y}^{(1)}} \int_{\sigma^2} g(\underline{\phi}) B(\underline{\phi}) (1/\sigma^2)^{(\ell+a)/2-1} \\
& \times e^{-\frac{1}{2\sigma^2} [b + \underline{y}^{(1)'} \underline{y}^{(1)}]} \prod_{i=\ell-q+1}^{\ell} \left\{ \int_{-\infty}^{\infty} h(\nu_i) d\nu_i \right\} d\sigma^2 d\underline{y}^{(1)} d\underline{\phi}. \tag{3.17}
\end{aligned}$$

We have shown in Appendix (E) that $\int_{-\infty}^{\infty} h(\nu_i) d\nu_i$, $i = \ell - q + 1, \dots, \ell$ is finite. Let $M(\underline{\phi})$ be the upper bound for $\prod_{i=\ell-q+1}^{\ell} \left\{ \int_{-\infty}^{\infty} h(\nu_i) d\nu_i \right\}$.

Now, integrating out σ^2 and noting that σ^2 has the inverse gamma distribution, we get

$$\begin{aligned}
F & < \int_{\underline{\phi}} \int_{\underline{y}^{(1)}} g(\underline{\phi}) M(\underline{\phi}) B(\underline{\phi}) \left\{ \int_0^{\infty} (1/\sigma^2)^{(\ell+a)/2-1} e^{-\frac{1}{2\sigma^2} [b + \underline{y}^{(1)'} \underline{y}^{(1)}]} d\sigma^2 \right\} d\underline{y}^{(1)} d\underline{\phi} \\
& = \int_{\underline{\phi}} \int_{\underline{y}^{(1)}} g(\underline{\phi}) M(\underline{\phi}) B(\underline{\phi}) \frac{\Gamma(\ell+a)/2}{[b + \underline{y}^{(1)'} \underline{y}^{(1)}]^{(\ell+a)/2}} d\underline{y}^{(1)} d\underline{\phi} \\
& = \int_{\underline{\phi}} \int_{\underline{y}^{(1)}} g(\underline{\phi}) M(\underline{\phi}) B(\underline{\phi}) \frac{\Gamma(\ell+a)/2}{[1 + \frac{\underline{y}^{(1)'} \underline{y}^{(1)}}{b}]^{(\ell+a)/2}} d\underline{y}^{(1)} d\underline{\phi}.
\end{aligned}$$

Here, the density of $\underline{y}^{(1)}$ (with dimension of $\underline{y}^{(1)}$, $t = \ell - q$) is a multivariate t-distribution with degree of freedom, $\vartheta = q + a$, $\underline{\mu} = \underline{0}$ and $\Sigma = (b/\nu)I$. Integrating out $\underline{y}^{(1)}$, we get

$$F < \int_{\underline{\phi}} g(\underline{\phi}) M(\underline{\phi}) B(\underline{\phi}) d\underline{\phi}.$$

Finally, because $g(\underline{\phi})$, $M(\underline{\phi})$, $B(\underline{\phi})$ are finite distinctly in $\epsilon \leq \phi_{sk} \leq 1 - \epsilon$, with small $\epsilon > 0$, $s = 1, \dots, S - 1$, $k = 0, \dots, p$, $\int_{\underline{\phi}} g(\underline{\phi}) M(\underline{\phi}) B(\underline{\phi}) d\underline{\phi}$ is finite. The mild condition we used is that $0 < \phi_{sk} < 1$.

3.1.2 Computation

First, we fitted (3.12) using the Gibbs sampler but we found that γ has high autocorrelation. So, we draw the samples using the composition rule as

$$\pi(\phi, \nu, \gamma, \sigma^2 | \underline{I}) = \pi(\phi, \nu, \sigma^2 | \underline{I})\pi(\gamma | \phi, \nu, \sigma^2, \underline{I}).$$

Integrating out γ (which has multivariate normal) from (3.9), we get

$$\begin{aligned} \pi(\phi, \nu, \sigma^2 | \underline{I}) &\propto \prod_{i=1}^{\ell} \left\{ \frac{\prod_{s=1}^{S-1} \prod_{k=0}^p \left(\frac{\phi_{sk}}{1-\phi_{sk}}\right)^{\sum_{j=1}^{n_i} x_{ijk} I_{ijs}} e^{\nu_i a_i}}{\prod_{j=1}^{n_i} \left[1 + e^{\nu_i} \sum_{s=1}^{S-1} \prod_{k=0}^p \left(\frac{\phi_{sk}}{1-\phi_{sk}}\right)^{x_{ijk}}\right]} \right\} \\ &\times (1/\sigma^2)^{(\ell+a)/2-1} \times e^{-\frac{1}{2\sigma^2} \left\{ b + \gamma_0' \frac{(z'z)}{\kappa+1} \gamma_0 + \nu' \left[I - \frac{\kappa}{\kappa+1} z(z'z)^{-1} z' \right] \nu - \frac{2}{\kappa+1} \gamma_0' z' \nu \right\}} \\ &\times \left\{ \prod_{s=1}^{S-1} \prod_{\kappa=0}^p \frac{1}{\phi_{sk}(1-\phi_{sk})} \right\}. \end{aligned} \quad (3.18)$$

The joint posterior density in (3.18) is complex, we use Markov chain Monte Carlo methods to fit it. Specifically, we use the grid method and adaptive rejection sampling to sample the parameters. First, we consider the conditional posterior distribution of ϕ

$$\begin{aligned} \pi(\phi | \nu, \sigma^2, \underline{I}) &\propto \frac{\prod_{s=1}^{S-1} \prod_{k=0}^p \left(\frac{\phi_{sk}}{1-\phi_{sk}}\right)^{\sum_{i=1}^{\ell} \sum_{j=1}^{n_i} x_{ijk} I_{ijs}} e^{\sum_{i=1}^{\ell} \nu_i a_i}}{\prod_{i=1}^{\ell} \prod_{j=1}^{n_i} \left[1 + e^{\nu_i} \sum_{s=1}^{S-1} \prod_{k=0}^p \left(\frac{\phi_{sk}}{1-\phi_{sk}}\right)^{x_{ijk}}\right]} \\ &\times \left\{ \prod_{s=1}^{S-1} \prod_{\kappa=0}^p \frac{1}{\phi_{sk}(1-\phi_{sk})} \right\}. \end{aligned} \quad (3.19)$$

We can write the denominator in a slightly simplified form to make ease for sampling

$$\begin{aligned} &\prod_{i=1}^{\ell} \prod_{j=1}^{n_i} \left[1 + e^{\nu_i} \sum_{s=1}^{S-1} \prod_{k=0}^p \left(\frac{\phi_{sk}}{1-\phi_{sk}}\right)^{x_{ijk}} \right] \\ &= \prod_{i=1}^{\ell} \prod_{j=1}^{n_i} \left[1 + e^{\nu_i} \left\{ \prod_{k=0}^p \left(\frac{\phi_{sk}}{1-\phi_{sk}}\right)^{x_{ijk}} + \sum_{s'=1, s' \neq s}^{S-1} \prod_{k=0}^p \left(\frac{\phi_{s'k}}{1-\phi_{s'k}}\right)^{x_{ijk}} \right\} \right] \\ &= \prod_{i=1}^{\ell} \prod_{j=1}^{n_i} \left[1 + e^{\nu_i} \left\{ \left(\frac{\phi_{sk}}{1-\phi_{sk}}\right)^{x_{ijk}} \prod_{k'=0, k' \neq k}^p \left(\frac{\phi_{sk'}}{1-\phi_{sk'}}\right)^{x_{ijk'}} + \sum_{s'=1, s' \neq s}^{S-1} \prod_{k=0}^p \left(\frac{\phi_{s'k}}{1-\phi_{s'k}}\right)^{x_{ijk}} \right\} \right]. \end{aligned}$$

We use an adaptive grid method (Ritter and Tanner, 1992) to draw a sample of each ϕ_{sk} from its univariate distribution $\pi(\phi_{sk} | \phi_{(sk)}, \nu, \sigma^2, \underline{I})$. We started by using 10 grids (i.e. we have divided the range of ϕ_{sk} , (0, 1), into 10 intervals of equal widths) to form an approximate

probability mass function of ϕ_{sk} , $s = 1, \dots, S-1$, $k = 0, \dots, p$ based on the evaluation of $\pi(\phi_{sk} | \phi_{(sk)\underline{z}}, \sigma^2, \underline{I})$ on a grid of point. Then, we determine an interval (a, b) of ϕ_{sk} of high mass. Typically (a, b) is much narrower than $(0, 1)$. We now take (a, b) as our new interval and stratify the range into 10 grids to approximate the probability density function by a probability mass function. Using this probability mass function, we draw ϕ_{sk} from (a, b) . We perform this for each ϕ_{sk} based on its conditional posterior densities.

Next, we consider the conditional posterior distribution of σ^2

$$\pi(\sigma^2 | \underline{\phi}, \underline{\nu}, \underline{I}) \propto (1/\sigma^2)^{(\ell+a)/2-1} \times e^{-\frac{1}{2\sigma^2} \left\{ b + \gamma_0' \frac{(z'z)}{\kappa+1} \gamma_0 + \underline{\nu}' \left[I - \frac{\kappa}{\kappa+1} z(z'z)^{-1} z' \right] \underline{\nu} - \frac{2}{\kappa+1} \gamma_0' z' \underline{\nu} \right\}}.$$

Therefore,

$$\sigma^2 | \underline{\phi}, \underline{\nu}, \underline{I} \sim \text{IGamma} \left\{ (\ell + a)/2, \frac{1}{2} \left[b + \gamma_0' \frac{(z'z)}{\kappa+1} \gamma_0 + \underline{\nu}' \left[I - \frac{\kappa}{\kappa+1} z(z'z)^{-1} z' \right] \underline{\nu} - \frac{2}{\kappa+1} \gamma_0' z' \underline{\nu} \right] \right\}. \quad (3.20)$$

We note that when $\kappa \rightarrow \infty$ (in which case the prior of γ is improper),

$$\sigma^2 | \underline{\phi}, \underline{\nu}, \underline{I} \sim \text{IGamma} \left\{ (\ell + a)/2, \frac{1}{2} (b + \underline{\nu}' [I - z(z'z)^{-1} z'] \underline{\nu}) \right\},$$

and similarly when $\kappa \rightarrow 0$,

$$\sigma^2 | \underline{\phi}, \underline{\nu}, \underline{I} \sim \text{IGamma} \left\{ (\ell + a)/2, \frac{1}{2} [b + \gamma_0' (z'z) \gamma_0 + \underline{\nu}' \underline{\nu} - 2\gamma_0' z' \underline{\nu}] \right\}.$$

The conditional posterior distribution of $\underline{\nu}$ is

$$\begin{aligned} \pi(\underline{\nu} | \underline{\phi}, \sigma^2, \underline{I}) &\propto \prod_{i=1}^{\ell} \left\{ \frac{e^{a_i \nu_i}}{\prod_{j=1}^{n_i} \left\{ 1 + e^{\nu_i} \sum_{s=1}^{S-1} \prod_{k=0}^p \left(\frac{\phi_{sk}}{1-\phi_{sk}} \right)^{x_{ijk}} \right\}} \right\} \\ &\times e^{-\frac{1}{2\sigma^2} \left\{ b + \gamma_0' \frac{(z'z)}{\kappa+1} \gamma_0 + \underline{\nu}' \left[I - \frac{\kappa}{\kappa+1} z(z'z)^{-1} z' \right] \underline{\nu} - \frac{2}{\kappa+1} \gamma_0' z' \underline{\nu} \right\}}. \end{aligned} \quad (3.21)$$

Again we note that when $\kappa \rightarrow \infty$,

$$\begin{aligned} \pi(\underline{\nu} | \underline{\phi}, \sigma^2, \underline{I}) &\propto \prod_{i=1}^{\ell} \left\{ \frac{e^{a_i \nu_i}}{\prod_{j=1}^{n_i} \left\{ 1 + e^{\nu_i} \sum_{s=1}^{S-1} \prod_{k=0}^p \left(\frac{\phi_{sk}}{1-\phi_{sk}} \right)^{x_{ijk}} \right\}} \right\} \\ &\times e^{-\frac{1}{2\sigma^2} \left\{ b + \underline{\nu}' [I - z(z'z)^{-1} z'] \underline{\nu} \right\}}, \end{aligned}$$

and when $\kappa \rightarrow 0$,

$$\pi(\underline{\nu} \mid \underline{\phi}, \sigma^2, \underline{I}) \propto \prod_{i=1}^{\ell} \left\{ \frac{e^{a_i \nu_i}}{\prod_{j=1}^{n_i} \left\{ 1 + e^{\nu_i} \sum_{s=1}^{S-1} \prod_{k=0}^p \left(\frac{\phi_{sk}}{1-\phi_{sk}} \right)^{x_{ijk}} \right\}} \right\} \\ \times e^{-\frac{1}{2\sigma^2} \left\{ b + \gamma_0' z' z \gamma_0 + \underline{\nu}' \underline{\nu} - 2 \gamma_0' z' \underline{\nu} \right\}}.$$

We use adaptive rejection sampling (ARS, Gilks and Wild, 1992) to draw ν_i , $i = 1, \dots, \ell$ from their conditional posterior densities. To apply the ARS, we need to show that the conditional posterior density of each ν_i in (3.21) is log-concave.

Let $P = [I - \frac{\kappa}{\kappa+1} z(z'z)^{-1}z']$ and $\underline{t} = \frac{1}{\kappa+1} z\gamma_0$, we have from (3.21)

$$\pi(\underline{\nu} \mid \underline{\phi}, \sigma^2, \underline{I}) = e^{-\frac{1}{2\sigma^2} \left\{ b + \underline{\nu}' P \underline{\nu} - 2 \underline{t}' \underline{\nu} \right\}} \times \prod_{i=1}^{\ell} h(\nu_i) \\ \propto e^{-\frac{1}{2\sigma^2} \left\{ \sum_{i=1}^{\ell} \sum_{i'=1}^{\ell} \nu_i \nu_{i'} p_{ii'} - 2 \sum_{i=1}^{\ell} t_i \nu_i \right\}} \times \prod_{i=1}^{\ell} h(\nu_i),$$

where $h(\nu_i) = \frac{e^{a_i \nu_i}}{\prod_{j=1}^{n_i} \{1 + e^{\nu_i} c_{ij}\}}$ is defined in (3.14) and $c_{ij} = \sum_{s=1}^{S-1} \prod_{k=0}^p \left(\frac{\phi_{sk}}{1-\phi_{sk}} \right)^{x_{ijk}}$ is defined in (3.13). Noting that P is symmetric, the function for fixed i is

$$\pi(\nu_i \mid \underline{\phi}, \sigma^2, \underline{I}) \propto e^{-\frac{1}{2\sigma^2} \left\{ (2\nu_i \sum_{i'=1}^{\ell} \nu_{i'} p_{ii'} - \nu_i^2 p_{ii}) - 2t_i \nu_i \right\}} \times h(\nu_i). \quad (3.22)$$

Lemma 3.1.2. $\pi(\nu_i \mid \underline{\phi}, \sigma^2, \underline{I})$ in (3.22) is log-concave.

Proof: Taking the logarithm of both sides of (3.22), we get

$$\Delta = \log[g(\nu_i)] = -\frac{1}{2\sigma^2} \left\{ (2\nu_i \sum_{i'=1}^{\ell} \nu_{i'} p_{ii'} - \nu_i^2 p_{ii}) - 2t_i \nu_i \right\} + \log\{h(\nu_i)\}.$$

The first and second order derivatives are

$$\Delta' = -\frac{1}{\sigma^2} \left\{ \sum_{i'=1}^{\ell} \nu_{i'} p_{ii'} - t_i \right\} + a_i - \sum_{j=1}^{n_i} \frac{c_{ij} e^{\nu_i}}{1 + c_{ij} e^{\nu_i}} \\ \Delta'' = -(1/\sigma^2) p_{ii} - \sum_{j=1}^{\ell} \frac{c_{ij} e^{\nu_i}}{(1 + c_{ij} e^{\nu_i})^2}.$$

Here, $p_{ii} \geq 0$ because P is a non-negative matrix. Therefore, $\Delta'' < 0$, so $\pi(\nu_i \mid \underline{\phi}, \sigma^2, \underline{I})$ is log-concave.

The sampling algorithm is executed by running the Gibbs sampler 5500 times, every time drawing a random deviate from (3.19), (3.20) and (3.21). We use a “burn in” of 500 iterates and take every fifth iterate thereafter from the remaining 5000 to make the auto-correlations among the iterates negligible. This provides a sample of $M = 1000$ from the posterior densities for the estimation. Finally, after drawing samples of $\hat{\phi}$, $\hat{\sigma}^2$ and $\hat{\nu}$, we draw $\hat{\gamma}$ from its conditional posterior density given the other parameters. The conditional posterior distribution for γ is

$$\gamma \mid \hat{\phi}, \hat{\nu}, \hat{\sigma}^2, \hat{I} \sim N_q \left\{ \frac{\kappa \hat{\gamma} + \gamma_0}{\kappa + 1}, \frac{\kappa}{\kappa + 1} \hat{\sigma}^2 \hat{\Delta} \right\}.$$

Again, we note that when $\kappa \rightarrow \infty$ (in which case the prior of γ is improper), $\gamma \mid \hat{\phi}, \hat{\nu}, \hat{\sigma}^2, \hat{I} \sim N_q \{ \hat{\gamma}, \hat{\sigma}^2 \hat{\Delta} \}$ and similarly as $\kappa \rightarrow 0$, $\gamma \mid \hat{\phi}, \hat{\nu}, \hat{\sigma}^2, \hat{I} \rightarrow \gamma_0$.

3.1.3 Assessing the Model Fit

Next, we need to check the fit of the model to the observed data. If the model fits well, then replicated data generated under the model should look similar to the observed data (Gelman, Carlin, Stern and Rubin 2004, Ch. 6). For this we would like to quantify the discrepancies between data and model, and assess whether they could have arisen by chance, under the model’s own assumption. In order to evaluate the fit of the model, we draw simulated values from the posterior predictive distribution of replicated data and compare these samples to the observed data. Any systematic difference between the simulations and the data indicate potential failings of the model (Gelman, Carlin, Stern and Rubin 2004, Ch. 6). Lack of fit of the data with respect to the posterior predictive distribution can be measured by the tail-area probability, or p-values, of the test quantity, and computed using posterior simulations of (θ, y^{rep}) .

We use a summary measure of fit, in particular the χ^2 discrepancy quantity, written in terms of univariate responses y_i as

$$\chi^2 \text{ discrepancy: } T(y, \theta) = \sum_i \frac{(y_i - E(y_i \mid \theta))^2}{\text{var}(y_i \mid \theta)},$$

where the summation is over the sample observations. The same summary measure can also

be calculated for posterior predictive distribution. Then, the posterior predictive p-value is

$$PPP = \Pr(T(y^{rep}, \theta) \geq T(y, \theta) \mid y).$$

In the context of our problem, we compute the χ^2 discrepancy quantity as

$$T(\underline{n}, \underline{a}) = \sum_{s=1}^S \frac{(n_s - E(n_s \mid \underline{a}))^2}{\text{var}(n_s \mid \underline{a})}, \quad (3.23)$$

where $n_s = \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} I_{ijs}$, $E(n_s \mid \underline{a}) = \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} E(I_{ijs}) = \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} a_{ijs}$ and $\text{var}(n_s \mid \underline{a}) = \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} a_{ijs}(1-a_{ijs})$. Note here that we use the total table to compute $T(\underline{n}, \underline{a})$ instead of using the cluster tables under which the corresponding quantity would be computed as $T(\underline{n}, \underline{a}) = \sum_{s=1}^S \sum_{i=1}^{\ell} \frac{(n_{is} - E(n_{is} \mid \underline{a}))^2}{\text{var}(n_{is} \mid \underline{a})}$. The reason for this is that we may have some cells with zero counts in cluster tables which leads to the problem of instability. We calculate $T(\underline{n}, \underline{a})$ for both observed and replicated data:

$$T(\underline{n}^{(obs)}, \underline{a}) = \sum_{s=1}^S \frac{(n_s^{(obs)} - \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} a_{ijs})^2}{\sum_{i=1}^{\ell} \sum_{j=1}^{n_i} a_{ijs}(1 - a_{ijs})}, \quad (3.24)$$

and

$$T(\underline{n}^{(rep)}, \underline{a}) = \sum_{s=1}^S \frac{(n_s^{(rep)} - \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} a_{ijs})^2}{\sum_{i=1}^{\ell} \sum_{j=1}^{n_i} a_{ijs}(1 - a_{ijs})}. \quad (3.25)$$

Replicate the data $M = 1000$ times and each time calculate $T(\underline{n}^{(rep)}, \underline{a})$. Then,

$$\begin{aligned} PPP &= \Pr\{T(\underline{n}^{(rep)}, \underline{a}) \geq T(\underline{n}^{(obs)}, \underline{a}) \mid \underline{n}^{(obs)}\} \\ &= \frac{\#[T(\underline{n}^{(rep)}, \underline{a}) \geq T(\underline{n}^{(obs)}, \underline{a}) \mid \underline{n}^{(obs)}]}{M}. \end{aligned}$$

The p-value close to 0 or 1 indicates that the observed pattern would be unlikely to be seen in replications of the data if the model were true with an extreme p-value, implying that the model cannot be expected to capture this aspect of the data. In order to address this problem of model failure, we need to improve the model in an appropriate way.

3.1.4 Surrogate Cluster Sample Without Covariates

We discuss how to obtain a surrogate sample without covariates. Upon fitting the cluster model with covariates in (3.1) and (3.3), we estimate the cell probabilities $(\hat{a}_{ijs}^{(h)}, i, \dots, \ell, j =$

$1, \dots, n_i, s = 1, \dots, S; h = 1, \dots, M)$ for each of the cluster tables by using the multinomial logistic regression. Let $\hat{a}_i^{(h)} = (\hat{a}_{i1}^{(h)}, \dots, \hat{a}_{iS}^{(h)})$, $h = 1, \dots, M$ denote the M estimates of the probability for i^{th} , $i = 1, \dots, \ell$, cluster table. Then, given these cell estimates of the cluster tables, we draw $\hat{n}_i^{(h)}$ as

$$\hat{n}_i^{(h)} \stackrel{ind}{\sim} \text{Multinomial}\{n_i, \hat{a}_i^{(h)}\}, \quad h = 1, \dots, M. \quad (3.26)$$

Note that, the sample data thus obtained represents the surrogate of the original cluster sample with covariates. The surrogate samples $\{\hat{n}_i^{(h)}, i = 1, \dots, \ell\}$ are now free of covariates.

3.2 Cluster Model Without Covariates

We use a hierarchical Bayesian model to convert $M = 1000$ cluster samples obtained in Section 3.1.4 to equivalent simple random samples. For this, instead of using the methodology developed in Chapter 2, we have developed a new methodology. We use the converted simple random sample to compute Bayes factor to make an inference about independence. We first describe the hierarchical Bayesian model and then show the computation.

3.2.1 Hierarchical Bayesian Model

For convenience, we drop the superscript h and use n_i for \hat{n}_i for the sample obtained in (3.26). Then, we assume

$$n_i \mid a_i \stackrel{ind}{\sim} \text{Multinomial}(n_i, a_i) \quad i = 1, \dots, \ell, \quad (3.27)$$

where $n_i = (n_{i1}, \dots, n_{iS})$, $n_i = \sum_{s=1}^S n_{is}$ and

$$a_{is} = \begin{cases} \frac{\pi_s e^{\nu_{is}}}{\pi_S + \sum_{s=1}^{S-1} \pi_s e^{\nu_{is}}}, & s = 1, \dots, S-1 \\ \frac{\pi_S}{\pi_S + \sum_{s=1}^{S-1} \pi_s e^{\nu_{is}}}, & s = S. \end{cases}$$

In (3.27) we have standard constraints $\{\sum_{s=1}^S a_{is} = 1, i = 1, \dots, \ell, \sum_{s=1}^S \pi_s = 1, a_{is} > 0, \pi_s > 0\}$. We want a test of independence based on the π_s .

A priori we assume

$$\nu_{is} \stackrel{iid}{\sim} N(0, \sigma^2), \quad i = 1, \dots, \ell, \quad s = 1, \dots, S-1; \quad (\sigma^2)^{-1} \sim \text{Gamma}(c, d), \quad (3.28)$$

where we choose $c = d = .001$. The likelihood function of the data is

$$\begin{aligned}
p(\underline{n} \mid \underline{\nu}, \underline{\pi}) &= \prod_{i=1}^{\ell} \left\{ n_i! \prod_{s=1}^S \frac{(a_{is})^{n_{is}}}{n_{is}!} \right\} \\
&\propto \prod_{i=1}^{\ell} \left\{ \prod_{s=1}^{S-1} \left[\frac{\pi_s e^{\nu_{is}}}{\pi_S + \sum_{s=1}^{S-1} \pi_s e^{\nu_{is}}} \right]^{n_{is}} \left[\frac{\pi_S}{\pi_S + \sum_{s=1}^{S-1} \pi_s e^{\nu_{is}}} \right]^{n_{iS}} \right\} \\
&= \prod_{i=1}^{\ell} \left\{ \frac{\prod_{s=1}^{S-1} (\pi_s e^{\nu_{is}})^{n_{is}} (1 - \sum_{s=1}^{S-1} \pi_s)^{n_{iS}}}{[\pi_S + \sum_{s=1}^{S-1} \pi_s e^{\nu_{is}}]^{n_i}} \right\}. \tag{3.29}
\end{aligned}$$

Combining the likelihood function in (3.29) with the prior densities in (3.28) via Bayes' theorem, the joint posterior density of $\underline{\pi}$, $\underline{\nu}$, and σ^2 given the surrogate cluster samples \underline{n} is

$$\begin{aligned}
\pi(\underline{\pi}, \underline{\nu}, \sigma^2 \mid \underline{n}) &\propto \prod_{i=1}^{\ell} \left\{ \frac{\prod_{s=1}^{S-1} (\pi_s e^{\nu_{is}})^{n_{is}} (1 - \sum_{s=1}^{S-1} \pi_s)^{n_{iS}}}{[\pi_S + \sum_{s=1}^{S-1} \pi_s e^{\nu_{is}}]^{n_i}} \right\} \\
&\times \prod_{i=1}^{\ell} \prod_{s=1}^{S-1} \left\{ (1/\sigma^2)^{1/2} e^{-\frac{1}{2\sigma^2} \nu_{is}^2} \right\} \times (1/\sigma^2)^{c-1} e^{-d(1/\sigma^2)}. \tag{3.30}
\end{aligned}$$

3.2.2 Computation

The joint posterior density in (3.30) is complex, we use Markov chain Monte Carlo methods to fit it. Specifically, we use the grid method and the Metropolis-Hastings sampler to sample the parameters. First, we consider the conditional posterior distribution of $\underline{\pi}$

$$\begin{aligned}
\pi(\underline{\pi} \mid \underline{\nu}, \sigma^2, \underline{n}) &\propto \prod_{i=1}^{\ell} \left\{ \frac{\prod_{s=1}^{S-1} (\pi_s e^{\nu_{is}})^{n_{is}} (1 - \sum_{s=1}^{S-1} \pi_s)^{n_{iS}}}{[\pi_S + \sum_{s=1}^{S-1} \pi_s e^{\nu_{is}}]^{n_i}} \right\} \\
&\propto \frac{(\prod_{s=1}^{S-1} \pi_s^{n_{\cdot s}}) (1 - \sum_{s=1}^{S-1} \pi_s)^{n_{\cdot S}}}{\prod_{i=1}^{\ell} [\pi_S + \sum_{s=1}^{S-1} \pi_s e^{\nu_{is}}]^{n_i}},
\end{aligned}$$

where $\sum_s^S \pi_s = 1$, $\pi_s \geq 0$, $s = 1, \dots, S$. Let $\underline{\pi}_{(s)}$ denote the vector of all components except π_s , then, we have

$$\pi(\pi_s \mid \underline{\pi}_{(s)}, \underline{\nu}, \sigma^2, \underline{n}) \propto \frac{\pi_s^{n_{\cdot s}} (1 - \pi_s - \sum_{s'=1, s' \neq s}^{S-1} \pi_{s'})^{n_{\cdot S}}}{\prod_{i=1}^{\ell} [\pi_S + \pi_s e^{\nu_{is}} + \sum_{s'=1, s' \neq s}^{S-1} \pi_{s'} e^{\nu_{is'}}]^{n_i}}, \tag{3.31}$$

$0 \leq \pi_s \leq 1 - \sum_{s'=1, s' \neq s}^{S-1} \pi_{s'}$. We use an adaptive grid method to draw a sample of π_s from its univariate distribution $\pi(\pi_s \mid \underline{\pi}_{(s)}, \underline{\nu}, \sigma^2, \underline{n})$. We started by using 10 grids (i.e. we have divided the range of π_s , $(0, 1 - \sum_{s'=1, s' \neq s}^{S-1} \pi_{s'})$, into 10 intervals of equal widths) to form an approximate probability mass function of π_s , $s = 1, \dots, S - 1$ based on the evaluation

of $\pi(\pi_s \mid \bar{\pi}_{(s)}, \nu, \sigma^2, \mathfrak{n})$ on a grid of points. Then, we determine an interval (a, b) of π_s of high mass. Typically (a, b) is much narrower than $(0, 1)$. We now take (a, b) as our new interval and stratify the range into 10 grids to approximate the probability density function by a probability mass function. Using this probability mass function, we draw π_s from $[a, b]$. Finally π_S is obtained from its conditional posterior density by taking $\pi_S = 1 - \sum_{s'=1, s' \neq s}^{S-1} \pi_{s'}$.

The conditional posterior density of ν is

$$\begin{aligned} \pi(\nu \mid \bar{\pi}, \sigma^2, \mathfrak{n}) &\propto \prod_{i=1}^{\ell} \left\{ \frac{\prod_{s=1}^{S-1} (e^{\nu_{is}})^{n_{is}}}{[\pi_S + \sum_{s=1}^{S-1} \pi_s e^{\nu_{is}}]^{n_i}} \right\} \times \prod_{i=1}^{\ell} \prod_{s=1}^{S-1} \left\{ e^{-\frac{1}{2\sigma^2} \nu_{is}^2} \right\} \\ &= \prod_{i=1}^{\ell} \left\{ \frac{\prod_{s=1}^{S-1} e^{(n_{is} \nu_{is} - \frac{1}{2\sigma^2} \nu_{is}^2)}}{[\pi_S + \sum_{s=1}^{S-1} \pi_s e^{\nu_{is}}]^{n_i}} \right\}. \end{aligned} \quad (3.32)$$

We use Metropolis-Hastings sampler to sample ν . For each i , we have

$$\pi(\nu_i \mid \bar{\pi}, \sigma^2, \mathfrak{n}) \propto \frac{\prod_{s=1}^{S-1} e^{(n_{is} \nu_{is} - \frac{1}{2\sigma^2} \nu_{is}^2)}}{[\pi_S + \sum_{s=1}^{S-1} \pi_s e^{\nu_{is}}]^{n_i}}. \quad (3.33)$$

The function of ν_i in (3.33) is unimodal, so using the mode Hessian approximation, we can approximate the density in (3.32) as

$$\nu_i \mid \gamma^2, \mathfrak{n} \sim N_p(\hat{\nu}_i, \gamma^2 \Sigma), \quad (3.34)$$

where $p = S - 1$. We obtain $\hat{\nu}_i$ and Σ by optimizing the function of ν_i in (3.33) using the Nelder-Mead algorithm. We take

$$\frac{\eta}{\gamma^2} \sim \chi_{\eta}^2. \quad (3.35)$$

We consider η as the tuning constant. Now combining (3.34) and (3.35) we get

$$\begin{aligned} \pi(\nu_i, \gamma^2 \mid \mathfrak{n}) &\propto (1/\gamma^2)^{p/2} e^{-\frac{1}{2\gamma^2} (\nu_i - \hat{\nu}_i)' \Sigma^{-1} (\nu_i - \hat{\nu}_i)} \times (\eta/\gamma^2)^{\eta/2-1} e^{-\frac{1}{2}(\eta/\gamma^2)} \\ &\propto (1/\gamma^2)^{(\eta+p)/2-1} e^{-\frac{1}{2\gamma^2} [(\nu_i - \hat{\nu}_i)' \Sigma^{-1} (\nu_i - \hat{\nu}_i) + \eta]}. \end{aligned} \quad (3.36)$$

Integrating out γ^2 (which has an inverse gamma distribution) from (3.36), we get

$$\pi(\nu_i \mid \mathfrak{n}) \propto \frac{1}{\left[1 + \frac{(\nu_i - \hat{\nu}_i)' \Sigma^{-1} (\nu_i - \hat{\nu}_i)}{\eta} \right]^{(\eta+p)/2}}$$

which is a multivariate t-distribution. Therefore,

$$\nu_i \mid \underline{n} \sim t_{(\eta+p)/2}(\hat{\nu}_i, \Sigma). \quad (3.37)$$

We consider this as our proposal density for ν_i to use Metropolis-Hastings algorithm.

Finally, the conditional posterior density of σ^2 is

$$\begin{aligned} \pi(\sigma^2 \mid \underline{\pi}, \underline{\nu}, \underline{n}) &\propto \prod_{i=1}^{\ell} \prod_{s=1}^{S-1} \left\{ (1/\sigma^2)^{1/2} e^{-\frac{1}{2\sigma^2} \nu_{is}^2} \right\} \times (1/\sigma^2)^{c-1} e^{-d(1/\sigma^2)} \\ &= (1/\sigma^2)^{[\ell(S-1)+2c/2]-1} e^{-\frac{1}{2\sigma^2} \{\sum_{i=1}^{\ell} \sum_{s=1}^{S-1} \nu_{is}^2 + 2d\}}. \end{aligned}$$

Therefore,

$$\sigma^{2-1} \sim \text{Gamma} \left[\{\ell(S-1) + 2c\}/2, \frac{1}{2} \left(\sum_{i=1}^{\ell} \sum_{s=1}^{S-1} \nu_{is}^2 + 2d \right) \right]. \quad (3.38)$$

The sampling algorithm is executed by running the Gibbs sampler 101 times, each time drawing a random deviate from (3.31), (3.37) and (3.38). Finally, we pick the last 101st sample value. We iterate this procedure for all $M = 1000$ cluster tables obtained in (3.26) after fitting the model in (3.2) and (3.3).

3.3 Bayes Factor

Having obtained samples of $\underline{\pi}$ from our cluster model defined by (3.27) and (3.28), we obtain simple random sample as

$$\underline{n} \sim \text{Multinomial}\{n, \underline{\pi}\},$$

Here, \underline{n} is surrogate data because the total table of cluster data without covariates has now been converted, and a model for simple random sampling is appropriate. To compute a value of the Bayes factor, we take

$$\underline{n} \sim \text{Multinomial}(n, \underline{\pi}), \quad \underline{\pi} \sim \text{Dirichlet}(\underline{v}), \quad (3.39)$$

where $v_s = .5, s = 1, \dots, S$, for Jeffreys' prior. In Section 1.3 we present the Bayes factor for a test of independence in the total table which is given in (1.5) as

$$BF = p_{\text{as}}(\underline{n})/p_{\text{nas}}(\underline{n}),$$

where $p_{\text{nas}}(n)$ and $p_{\text{as}}(n)$ are, respectively, the marginal likelihoods under the models with no association (nas) and association (as). We repeat this for all $M = 1000$ tables and compute the Bayes factor each time, yielding 1000 Bayes factor values. Then, we can find the modal Bayes factor value that will be used to make an inference about the test of independence. We can also obtain other summaries of the Bayes factor.

3.4 Applications

In this section, we present an illustrative example using TIMSS 2007 fourth grade US data. We described the data in Chapter 1 and the cluster tables are presented in Appendix F. We consider two variables: mathematics test score (below average and above average) and science test scores (below average and above average). We study the association between mathematics test scores (MTS) and science test scores (STS) by community. This creates six examples in all.

In Table 3.1, we present the total tables for the six examples (E1-E6 for MTS versus STS in each of the six communities). The number (ℓ) of clusters changes considerably over the communities, as does the number of observations. The intracluster correlations are small and they do not vary too much over examples. The design effects (DEFs), obtained from Brier’s model, are larger than one but not so big. The observed counts in the total tables are large in all examples.

We have five student (unit) level covariates: (i) Sex (X_1), (ii) How often do you speak English at home? (X_2), (iii) Index of self confidence learning math (X_3), (iv) Index of self confidence learning science (X_4) and (v) Race (X_5), and three school (cluster) level covariates: (a) Approximately what percentage of students in school come from economically affluent homes? (Z_1), (b) Percent of free lunch-categorized (Z_2) and (c) Total school enrollment in all grades (Z_3). We do not present the data on covariates.

From the preliminary analysis, we have found that the unit level covariates X_1 and X_5 are insignificant across most of the cells for all examples, and similarly the cluster level covariates Z_2 and Z_3 are also insignificant over examples. However, the remaining covariates (X_2 , X_3 , X_4 and Z_1), although they are not significant uniformly, we keep them to make the

set of covariates common across all the examples. We fit the cluster model with covariates by using the final set of covariates. The posterior estimates of regression coefficients together with the 95% credible intervals are presented in Table 3.2 by examples.

In order to assess the fit of our model, we compute the chi-squared discrepancy measure for the observed data and for the posterior simulations. We have obtained the posterior predictive p-values (PPP) for examples E1-E6 to be 0.467, 0.470, 0.489, 0.466, 0.447 and 0.500 respectively which are also presented in Table 3.2. These values indicate that in all examples the model fits the observed data very well. But note how they all cluster near 0.50, see Hjort, Dahl and Steinbakk (2006).

In Table 3.3, we present summaries of the log-Bayes factor as developed in Section 3.3. The modal log-Bayes factor values for examples E1-E6 are respectively 77.28, 76.39, 97.78, 90.61, 107.76 and 41.27. According to Kass and Raftery (1995), these values indicate very strong evidence for dependence between mathematics and science scores.

We have also computed the posterior predictive p-value to assess the fit of the second model, presented in Section 3.2. These are obtained by computing the chi-squared discrepancy measure $T(\underline{n}, \underline{a})$ as in (3.23) where now $n_s = \sum_{i=1}^{\ell} n_{is}$, $E(n_s | \underline{a}) = \sum_{i=1}^{\ell} a_{is}$, and $\text{var}(n_s | \underline{a}) = \sum_{i=1}^{\ell} a_{is}(1 - a_{is})$. We compute $T(\underline{n}, \underline{a})$ for both observed data and replicated data. The posterior predictive p-values are 0.465, 0.469, 0.504, 0.448, 0.471 and 0.501 for examples E1-E6 respectively. These values show that the cluster model (3.27)-(3.28) is also fitting well in all examples but again the PPP values cluster near 0.50.

We have studied the effect of the student level (X_2 , X_3 and X_4) and school level (Z_1) covariates in the test of independence. For this, we started fitting the cluster model with the intercept only. Thereafter, we keep adding one student level covariate at a time and finally, we add the school level covariate. In Table 3.5, we present the results of the fit of these different models. We see that the model fit is good for each examples. The modal log-Bayes factor value is always large, with or without covariates. This suggests that the effect of the covariates is small and it is true for all examples. The reason for this may be that the two categorical variables (mathematics and science scores) are highly correlated. Inclusion of the school covariate changes the log-Bayes factors to some extent in all examples except E1 but

they are still large. For example in E1, the modal log-Bayes factor with the covariates X_2, X_3 and X_4 is 76.59, and with X_2, X_3, X_4 and Z_1 it is 77.28. Also in E2, the corresponding values are 87.74 and 76.39. Note here that the model with only the intercept term is the cluster model without covariates.

In order to make a meaningful comparison, we perform a test of independence with covariates but no random effect. This is equivalent to the test of independence with covariates under simple random sampling. The results of this study are presented in Table 3.6. The log-Bayes factor values are still large, showing a strong association between mathematics and science test scores. This result also shows that the effect of clustering is small. We also studied the effect of covariates, and we see that these do not have a substantial impact on the test.

We have computed the Bayes factor by treating the observed cluster sample as a simple random sample. For this we simply combine the cluster tables, ignoring the covariates. We have obtained the modal log-Bayes factor values for examples E1-E6 to be respectively 73.24, 90.28, 105.80, 96.40, 113.12 and 47.63. These are close to the corresponding values for cluster model with covariates presented in Table 3.3 as one might expect.

Looking at these different cases, there appears to be a very strong dependence between mathematics score and science score with and without covariates. To further understand and investigate this result, we perform a simulation study below in Section 3.5.

We have also computed the posterior design effects for the individual cells. These design effects are computed as the ratio of diagonal elements of the posterior variance of $\pi = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$ under the multinomial logistic regression random effect model specified by (3.1) and (3.2) and the corresponding posterior variance under the model for simple random sampling described below

$$I_{ij} \stackrel{ind}{\sim} \text{Multinomial}(1, \underline{a}_{ij}), \quad i = 1, \dots, \ell, \quad j = 1, \dots, n_i, \quad (3.40)$$

where

$$a_{ijs} = \begin{cases} \frac{e^{\underline{\beta}_s' \underline{x}_{ij} + \underline{\gamma}' z_i}}{1 + \sum_{s=1}^{S-1} e^{\underline{\beta}_s' \underline{x}_{ij} + \underline{\gamma}' z_i}}, & s = 1, \dots, S-1 \\ \frac{1}{1 + \sum_{s=1}^{S-1} e^{\underline{\beta}_s' \underline{x}_{ij} + \underline{\gamma}' z_i}}, & s = S, \end{cases}$$

and a priori

$$\pi(\underline{\beta}) \propto k_1 \text{ and } \pi(\underline{\gamma}) \propto k_2, \quad (3.41)$$

where we take $k_1 = 1$ and $k_2 = 1$. After fitting the two models separately, we estimate $\pi_s = \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} a_{ijs}$, $s = 1, \dots, S$ using each $M = 1000$ sample of $\underline{\beta}_s$ and $\underline{\gamma}$. Then, we compute the covariance matrix of $\underline{\pi}$. The design effects are presented in Table 3.4. The design effects vary across the cell but not much larger than 1. The average design effects are 2.51, 1.66, 1.31, 1.34, 1.94 and 2.11 for E1-E6 respectively, some of which are very different than the analogue in Table (3.1) obtained from Brier's model. For example in E2 it is 4.25 using Brier's model and is 1.66 under our model. This is expected because we did not use covariates when fitting Brier's model, whereas under our model, we have incorporated the covariates.

In Table 3.7, we study the issue of sensitivity of the specification of the parameters a and b used in the prior for σ^2 in (3.3) to the Bayesian test of independence. We varied a and b to be .001, .01, .1 and 1.0. The evidence against independence does not change markedly. This is true in all six examples.

3.5 Simulation Study

We have performed a simulation study to help understand the tests further. We consider three factors: dependence between the two categorical variables (weak, medium, strong), number of clusters (low, medium) and intracluster correlation (very small, small, moderate).

Corresponding to the four cells (2×2 categorical table), let $\psi_s = 1$, $s = 2, 3$ the (off-diagonal cells) and $\psi_s = \text{ind}$, $s = 1, 4$ (diagonal cells), where 'ind' is to be specified. The cell probabilities are $\psi_s / \sum_{s=1}^S \psi_s$, $s = 1, \dots, S$. When the ψ_s are roughly the same (ind=1), there will be independence and when the diagonal ψ_s are larger than 1, there will be dependence (ind=2). For a 2×2 table with large cell counts, if the diagonal probabilities are twice the off diagonals, there will be strong dependence (ind=2). With an intracluster correlation of ρ , we set $\alpha_s = \{(1 - \rho)/\rho\} \psi_s / \sum_{s=1}^S \psi_s$, $s = 1, \dots, S$. For $i = 1, \dots, \ell$, we generate $\pi_i \stackrel{iid}{\sim} \text{Dirichlet}(\underline{\alpha})$ to get the cell probabilities for the cluster tables without covariates. We

get the corresponding probabilities for the cluster tables with covariates as

$$\tilde{\pi}_{ijs} = \begin{cases} \frac{\pi_{is} e^{\beta'_s x_{ij} + \gamma' z_i}}{\pi_{iS} + \sum_{s=1}^{S-1} \pi_{is} e^{\beta'_s x_{ij} + \gamma' z_i}}, & s = 1, \dots, S-1 \\ \frac{\pi_{is}}{\pi_{iS} + \sum_{s=1}^{S-1} \pi_{is} e^{\beta'_s x_{ij} + \gamma' z_i}}, & s = S. \end{cases}$$

Note here that we use the student level and school level covariates from the observed data but generate the new data. We divide the total number of observations into the clusters with sizes, n_i , $i = 1, \dots, \ell$, based on a multinomial distribution with equal cell probabilities. Finally, the cluster tables are generated independently by drawing an indicator I_{ij} from multinomial distributions with cell probabilities $\tilde{\pi}_{ij}$, so that total count is n_i . In order to construct an example we choose the covariates X_3 and Z_2 from example E2 because these are the most significant covariates. Accordingly we use the cluster size $\ell = 27$ (with sample size 837) that corresponds to E2. We choose $\text{ind} = 1.0, 1.2, 1.4$, $\rho = .01, .10, .30$ and $\ell = 27, 54$. In order to obtain the covariates for $\ell = 54$ cluster, we simply stacked the covariates from E2 twice. The sample size for $\ell = 54$ clusters would be 1674.

Thus, there are eighteen ($3 \times 3 \times 2$) design points, and 100 cluster samples are generated at each design point. We perform our computations exactly as for TIMSS 2007 fourth grade US data and obtain the Bayes factors from our model. We fit (i) a model with covariates (MWC) and (ii) a model without covariates (MWOC) for the same data. We ‘average’ various quantities over the 100 replications at each design point. For example, in Table 3.8 the mode is the average of the 100 modes.

In Table 3.8 and Table 3.9, we present numerical summaries from the simulation study. We see that mostly, the log-Bayes factor values increase as ρ increases. This is because usually high intraclass correlation dampens the effects. Comparing the corresponding modal values for MWC and MWOC, there are some notable differences as marked with “†” in the tables. For example, when $\text{ind}=1.0$, $\rho = .30$ and $\ell = 27$ the modes of the log-Bayes factors are -0.22 and 3.53 respectively for MWC and MWOC. These two values provides different inferences about the test of independence. According to Kass and Raftery (1995), the first one is the evidence for independence and the second one is the ‘strong’ evidence against the independence. Similarly, with $(\text{ind} = 1.2, \rho = .10, \ell = 27)$, the log-factor values are 0.20 and -0.62 . We also see these kinds of differences for $(\text{ind} = 1.2, \rho = .30, \ell = 27)$, $(\text{ind} = 1.4, \rho =$

.01, $\ell = 27$), (ind = 1.2, $\rho = .01$, $\ell = 54$) and (ind = 1.2, $\rho = .10$, $\ell = 54$). The interquartile ranges of the log-Bayes factor gets narrower as the dependence structure increases (i.e., as ind increases). Also as the structure gets more dependence, the log-Bayes factor increases showing that there are no differences in an inference for the test of independence between two models, as we found in real data.

In Figure 3.1, we plot the empirical distributions of the log-Bayes factors when fitting the model with covariates for each of the design points. Of the three plots, the first corresponds to ind=1.0 and has six curves corresponding to six design points ($\rho = .01$, $\ell = 27$), ($\rho = .10$, $\ell = 27$), ($\rho = .30$, $\ell = 27$), ($\rho = .01$, $\ell = 54$), ($\rho = .10$, $\ell = 54$), ($\rho = .30$, $\ell = 54$). Similarly, the other two plots correspond to ind=1.2 and ind=1.4. Comparing these plots we see that distribution of the log-Bayes factor has least variation when ind=1.0 (independent), and has largest with ind=1.4 (moderately dependent). We have also obtained the these plots when fitting the model without covariates. They appear similar to the one with covariates.

3.6 Power Function

We calculate the ‘power’ of our statistical test for the cluster model with covariates. In non-Bayesian statistics, the power of a statistical test is the probability that the test will reject the null hypothesis when the null hypothesis is false. To describe departures from the null hypothesis of independence, we consider a mixture distribution under the alternative hypothesis.

We assume that the distribution under the alternative hypothesis is multinomial with cell probabilities $\tilde{\pi}_{ijk_u} = w\pi_{ijk_u} + (1 - w)\pi_{k \cdot} \pi_{\cdot u}$, $i = 1, \dots, \ell$, $j = 1, \dots, n_i$, $k = 1, 2$, $u = 1, 2$; $\pi_{k \cdot} = \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \sum_{u=1}^2 \pi_{ijk_u}$, $\pi_{\cdot u} = \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \sum_{k=1}^2 \pi_{ijk_u}$. Note here that π_{ijk_u} is arbitrary, the more important is w . So, for π_{ijk_u} , we use the estimate of a_{ijk_u} , which are obtained after fitting the cluster model with covariates to the observed data. Here, $w = 0$ corresponds to the null hypothesis distribution and $w = 1$ corresponds to the alternative hypothesis of dependence. Therefore, values of w close to zero give local alternatives, and larger values of w give larger departure from the null hypothesis.

Letting bf_{α} denote the critical value of an upper-tailed test of size α , the power function

is given by

$$Pr\{BF > bf_\alpha \mid I_{ij} \sim \text{Multinomial}(1, \tilde{\pi}_{ij}), i = 1, \dots, \ell, j = 1, \dots, n_i\},$$

where $\sum_{i=1}^{\ell} n_i = n$. Here, BF is the Bayes factor test statistic used in our test, and ℓ and n_i are the number of clusters and sample size for Example E1 in observed data. The data is generated as

$$I_{ij} \sim \text{Multinomial}(1, \tilde{\pi}_{ij}), i = 1, \dots, \ell, j = 1, \dots, n_i, \quad (3.42)$$

and then the cluster models (3.1) and (3.3) are fitted to this data in a similar way to how they were fit to the TIMMS 2007 fourth grade US data and compute the Bayes factors.

The critical value is obtained by taking the $100(1 - \alpha)^{\text{th}}$ percentile point of the test statistics from the data generated under the null hypothesis. Generating data under the null hypothesis here is equivalent to generating data from (3.42) when $w = 0$. We generate $M = 1000$ data sets and fit the cluster models (3.1) and (3.3) thereby obtaining the Bayes factor for each set. We then obtain the critical value using the distribution of the Bayes factor. Since the power function in (3.42) is a function of $w \in [0, 1]$, we vary w in this range. For each w , we generated $M = 1000$ data sets and computed the Bayes factor (denoted by BF), thereby having $M = 1000$ Bayes factor values. Then, we obtain the proportion of values of the test statistics exceeding its critical value as

$$\text{p-value} = \frac{1}{M} \sum_{m=1}^M I\{BF > bf_\alpha\},$$

which is the power.

We have plotted the estimated power function in Figure 3.2. We see that the power function increases rapidly as w increases from 0 to 0.40, attaining power = 1.0 as w takes the value around 0.50, thereby showing that this is a reasonable test.

3.7 Concluding Remarks

We have proposed a method to the test of independence with covariates in an $r \times c$ contingency table obtained from a two-stage cluster sampling design with simple random sampling at both stages. We have used a hierarchical Bayesian model and a Markov chain Monte Carlo

method to fit it. We use the Bayes factor to make an inference about independence. For the real data (TIMSS 2007) we have found that there is very strong dependence between two categorical variables both with and without covariates. To further investigate and understand the findings, we have performed a simulation study where we generated the data under (i) the dependent model and (ii) the independent model. We then fit both the models with and without covariates. We have found that there are some noticeable differences between the test of independence from two models. We have also found that as the structure of the model from which we generate the data gets more dependent, there is no difference in the test of independence between two models.

Table 3.1: Features of the total table for each of the six examples

	n	ℓ	ρ	Def	(1,1)	(1,2)	(2,1)	(2,2)
E1	781	26	.078	3.52	229	116	101	335
E2	837	27	.100	4.25	239	101	117	380
E3	744	26	.090	3.71	239	88	85	332
E4	1467	49	.045	2.35	268	235	168	796
E5	1675	59	.064	2.84	412	305	205	753
E6	575	25	.056	2.41	141	112	53	269

NOTE: These are all 2×2 contingency tables; ρ is the intracluster correlation; Def stands for design effect calculated using Brier's model.

Table 3.2: Posterior estimate of the parameters under multinomial logistic regression

Examples	Cells	Covariates	PM	PSD	NSE	Interval
E1	1	Intercept	0.26	0.30	0.04	(-0.30, 0.85)
	1	X_2	0.06	0.06	0.00	(-0.08, 0.16)
	1	X_3	1.09	0.17	0.01	(0.77, 1.41)
	1	X_4	0.27	0.17	0.01	(-0.04, 0.61)
	2	Intercept	-0.95	0.32	0.04	(-1.53, -0.26)
	2	X_2	-0.17	0.07	0.01	(-0.32, -0.04)
	2	X_3	0.81	0.20	0.01	(0.37, 1.15)
	2	X_4	0.15	0.20	0.01	(-0.24, 0.54)
	3	Intercept	-1.15	0.34	0.04	(-1.81, -0.52)
	3	X_2	0.05	0.07	0.00	(-0.08, 0.20)
	3	X_3	-0.14	0.25	0.02	(-0.61, 0.35)
	3	X_4	0.58	0.20	0.01	(0.19, 0.93)
			Z_2	0.46	0.12	0.01
PPP=0.467						
E2	1	Intercept	0.72	0.22	0.02	(0.31, 1.14)
	1	X_2	0.14	0.06	0.00	(0.02, 0.24)
	1	X_3	1.32	0.17	0.01	(1.00, 1.65)
	1	X_4	0.37	0.15	0.01	(0.07, 0.63)
	2	Intercept	-0.68	0.26	0.02	(-1.19, -0.18)
	2	X_2	-0.03	0.08	0.00	(-0.17, 0.13)
	2	X_3	0.83	0.21	0.01	(0.45, 1.25)
	2	X_4	0.21	0.19	0.01	(-0.16, 0.57)
	3	Intercept	-0.56	0.25	0.02	(-1.04, -0.09)
	3	X_2	0.09	0.07	0.00	(-0.05, 0.23)
	3	X_3	0.51	0.21	0.01	(.09, 0.89)
	3	X_4	0.17	0.18	0.01	(-0.20, 0.53)
			Z_2	0.66	0.10	0.00
PPP=0.470						
E3	1	Intercept	1.22	0.26	0.03	(0.76, 1.79)
	1	X_2	0.31	0.07	0.01	(0.16, 0.46)
	1	X_3	1.22	0.18	0.01	(0.90, 1.60)
	1	X_4	0.39	0.16	0.01	(0.12, 0.72)
	2	Intercept	-0.42	0.31	0.03	(-0.98, 0.26)
	2	X_2	0.08	0.09	0.01	(-0.10, 0.27)
	2	X_3	1.32	0.21	0.01	(0.93, 1.71)
	2	X_4	-0.05	0.23	0.01	(-0.52, 0.39)
	3	Intercept	-0.48	0.30	0.03	(-1.04, 0.12)
	3	X_2	0.17	0.08	0.01	(0.00, 0.31)
	3	X_3	0.47	0.25	0.02	(-0.04, 0.91)
	3	X_4	0.39	0.19	0.01	(0.03, 0.75)
			Z_2	0.61	0.08	0.00
PPP=0.489						

<hr/>						
E4	1	Intercept	0.33	0.17	0.02	(0.01, 0.66)
	1	X_2	0.23	0.05	0.00	(0.12, 0.32)
	1	X_3	1.02	0.12	0.01	(0.78, 1.24)
	1	X_4	0.38	0.13	0.01	(0.16, 0.64)
	2	Intercept	-0.39	0.18	0.02	(-0.75, -0.05)
	2	X_2	0.11	0.05	0.00	(0.00, 0.20)
	2	X_3	0.93	0.12	0.01	(0.71, 1.17)
	2	X_4	-0.15	0.15	0.01	(-0.39, 0.16)
	3	Intercept	-0.45	0.20	0.02	(-0.83, -0.09)
	3	X_2	0.21	0.06	0.00	(0.09, 0.31)
	3	X_3	0.44	0.15	0.01	(0.18, 0.75)
	3	X_4	0.36	0.14	0.01	(0.09, 0.64)
		Z_2	0.42	0.07	0.003	(0.28,0.55)
PPP=0.466						
<hr/>						
E5	1	Intercept	0.93	0.19	0.02	(0.54, 1.27)
	1	X_2	0.26	0.06	0.01	(0.16, 0.37)
	1	X_3	0.93	0.11	0.01	(0.75, 1.15)
	1	X_4	0.71	0.11	0.01	(0.51, 0.93)
	2	Intercept	-0.22	0.23	0.03	(-0.68, 0.19)
	2	X_2	0.00	0.07	0.01	(-0.13, 0.12)
	2	X_3	0.93	0.11	0.01	(0.71, 1.14)
	2	X_4	0.25	0.13	0.01	(0.00, 0.48)
	3	Intercept	-0.27	0.22	0.02	(-0.71, 0.16)
	3	X_2	0.25	0.06	0.01	(0.12, 0.36)
	3	X_3	0.16	0.15	0.01	(-0.13, 0.46)
	3	X_4	0.55	0.13	0.01	(0.31, 0.80)
		Z_2	0.47	0.09	0.003	(0.29,0.67)
PPP=0.447						
<hr/>						
E6	1	Intercept	1.07	0.32	0.04	(0.46, 1.72)
	1	X_2	0.23	0.09	0.01	(0.05, 0.40)
	1	X_3	1.03	0.18	0.01	(0.68, 1.38)
	1	x_4	0.99	0.20	0.01	(0.63, 1.37)
	2	Intercept	-0.37	0.38	0.04	(-1.13, 0.42)
	2	X_2	-0.09	0.12	0.01	(-0.32, 0.12)
	2	X_3	0.68	0.19	0.01	(0.28, 1.03)
	2	X_4	0.42	0.24	0.02	(-0.07, 0.88)
	3	Intercept	-0.67	0.44	0.05	(-1.48, 0.24)
	3	X_2	0.02	0.13	0.01	(-0.22, 0.29)
	3	X_3	0.43	0.26	0.02	(-0.05, 0.95)
	3	X_4	0.95	0.25	0.01	(0.49, 1.41)
		Z_2	0.17	0.01	0.036	(0.04,0.67)
PPP=0.500						
<hr/>						

NOTE: (i) Posterior means (PM), posterior standard deviations (PSD), numerical standard errors (NSE) and 95% credible intervals for the regression coefficients
(ii) PPP denotes the posterior predictive p-value.

Table 3.3: Summary of the log-Bayes factor

Examples	Min	log-Bayes Factor			Max	Mode
		Q1	Q2	Q3		
E1	7.51	55.43	77.85	101.60	183.70	77.28
E2	10.59	70.32	95.70	122.50	230.40	76.39
E3	19.42	84.66	109.60	137.00	285.70	97.78
E4	7.39	73.76	99.47	128.10	279.00	90.61
E5	6.00	95.41	122.90	156.80	263.60	107.76
E6	3.34	33.79	48.96	66.66	142.60	41.27

Table 3.4: Bayesian design effects for each cell by example

Cell	E1	E2	E3	E4	E5	E6
(1,1)	2.07	1.53	1.28	1.16	1.58	1.58
(1,2)	1.44	1.17	1.01	1.23	1.42	1.80
(2,1)	1.52	1.24	1.19	1.08	1.38	1.19
(2,2)	5.03	2.73	1.78	1.90	3.40	3.87
Avg Def	2.51	1.66	1.31	1.34	1.94	2.11
ESS	311	504	568	1091	863	273

NOTE: The cells are (j, k) , $j, k = 1, 2$. Avg Def stands for average design effect. ESS stands for the effective sample size and it is the sum of the cell counts divided by the design effects, taken for the total table.

Table 3.5: Study of the effects of covariates on the test of independence

Examples	Covariates	PPP	log-Bayes Factor					
			Min	Q1	Q2	Q3	Max	Mode
E1 n=781	Intercept	.479	.10	57.49	78.11	101.10	243.60	73.89
	X_2	.478	1.79	54.83	75.53	100.50	255.10	67.79
	X_2X_3	.505	2.65	55.63	74.58	96.94	179.60	71.80
	$X_2X_3X_4$.491	6.08	54.16	76.49	99.29	206.30	76.59
	$X_2X_3X_4Z_1$.467	7.51	55.43	77.85	101.60	183.70	77.28
E2 n=837	Intercept	.471	.22	68.96	92.88	120.10	220.10	88.61
	X_2	.495	6.60	68.52	95.34	123.30	216.50	95.50
	X_2X_3	.472	11.60	73.25	97.24	125.70	225.60	91.59
	$X_2X_3X_4$.501	3.57	72.72	96.76	123.30	276.10	87.74
	$X_2X_3X_4Z_1$.470	10.59	70.32	95.70	122.50	230.40	76.39
E3 n=744	Intercept	.472	19.03	86.59	112.30	139.20	276.50	108.12
	X_2	.502	15.64	89.14	110.90	139.30	259.20	106.16
	X_2X_3	.467	12.63	85.13	111.20	139.10	267.30	91.67
	$X_2X_3X_4$.476	22.89	84.32	110.50	139.30	228.90	105.87
	$X_2X_3X_4Z_1$.489	19.42	84.66	109.60	137.00	285.70	97.78
E4 n=1467	Intercept	.467	6.07	78.44	102.10	129.20	243.70	96.67
	X_2	.449	7.87	73.52	100.80	129.70	276.60	98.97
	X_2X_3	.470	22.00	77.44	100.50	127.70	251.50	91.19
	$X_2X_3X_4$.447	12.21	73.47	98.52	125.00	274.60	79.79
	$X_2X_3X_4Z_1$.466	7.39	73.76	99.47	128.10	279.00	90.61
E5 n=1675	Intercept	.454	29.91	91.94	119.90	148.50	353.30	117.37
	X_2	.441	25.67	93.79	122.60	156.80	277.20	112.48
	X_2X_3	.470	19.03	94.00	119.40	149.40	280.10	114.78
	$X_2X_3X_4$.483	25.88	94.58	121.60	149.80	267.20	115.64
	$X_2X_3X_4Z_1$.447	6.00	95.41	122.90	156.80	263.60	107.76
E6 n=575	Intercept	.502	1.72	35.60	49.22	67.57	151.70	44.41
	X_2	.513	-0.80	34.05	48.56	65.98	138.80	42.03
	X_2X_3	.476	3.00	36.18	51.44	69.36	143.40	47.69
	$X_2X_3X_4$.493	1.15	36.32	51.17	67.16	144.90	48.83
	$X_2X_3X_4Z_1$.500	3.34	33.79	48.96	66.66	142.60	41.27

NOTE: PPP denotes the posterior predictive p-value, and n denotes the sample size.

Table 3.6: Study the effects of covariates under simple random sampling

Examples	Covariates	PPP	log-Bayes Factor					Max	Mode
			Min	Q1	Q2	Q3			
n=781	E1 Intercept	.494	19.03	57.50	72.23	89.75	162.30	68.54	
	X_2	.451	7.965	57.72	74.52	91.08	163.70	73.62	
	X_2X_3	.491	16.24	56.92	73.05	90.33	162.50	73.75	
	$X_2X_3X_4$.458	15.56	56.30	72.21	90.81	164.60	66.95	
	$X_2X_3X_4Z_1$.440	9.50	56.61	76.13	97.72	179.60	65.52	
n=837	E2 Intercept	.445	15.54	72.67	90.57	108.60	198.00	92.23	
	X_2	.465	20.35	73.13	89.93	109.50	199.40	84.00	
	X_2X_3	.461	9.383	71.30	87.60	110.20	173.70	81.41	
	$X_2X_3X_4$.496	16.92	72.89	89.05	109.20	218.80	82.89	
	$X_2X_3X_4Z_1$.474	12.15	71.00	93.71	121.30	272.30	81.17	
n=744	E3 Intercept	.471	40.41	86.16	104.90	124.30	212.20	104.38	
	X_2	.476	20.81	87.47	107.00	126.80	233.90	106.26	
	X_2X_3	.478	32.79	87.22	104.10	125.20	203.20	98.45	
	$X_2X_3X_4$.450	43.01	85.83	106.40	126.40	198.00	108.45	
	$X_2X_3X_4Z_1$.460	28.78	87.23	110.90	134.80	245.10	108.36	
n=1467	E4 Intercept	.444	26.41	75.43	95.77	120.10	269.90	92.04	
	X_2	.437	18.81	75.42	97.26	119.50	241.30	99.95	
	X_2X_3	.441	19.06	77.10	97.48	120.20	215.70	93.27	
	$X_2X_3X_4$.440	13.24	74.93	96.56	117.80	207.10	97.51	
	$X_2X_3X_4Z_1$.443	15.12	75.72	100.20	124.60	240.10	102.19	
n=1675	E5 Intercept	.439	33.44	87.06	111.20	135.60	236.40	110.940	
	X_2	.398	34.18	90.93	112.30	137.20	239.40	102.00	
	X_2X_3	.411	24.15	91.54	112.90	137.20	250.80	109.46	
	$X_2X_3X_4$.446	30.14	88.21	111.30	137.40	238.90	107.68	
	$X_2X_3X_4Z_1$.448	28.25	92.32	118.10	121.10	145.80	113.94	
n=575	E6 Intercept	.478	5.15	35.37	47.76	60.82	127.10	44.80	
	X_2	.468	0.90	34.40	47.26	61.90	168.00	37.02	
	X_2X_3	.463	4.31	35.95	48.60	62.49	111.10	46.89	
	$X_2X_3X_4$.444	4.68	34.89	47.07	62.58	133.90	42.36	
	$X_2X_3X_4Z_1$.470	5.30	35.83	50.17	67.04	138.80	44.94	

NOTE: PPP denotes the posterior predictive p-value, and n denotes the sample size.

Table 3.7: Sensitivity analysis of the log-Bayes factor with respect to a and b in the prior of σ^2 , by examples

Examples	$a(= b)$	<u>MTS vs. STS</u>			
		.001	.01	.1	1.0
E1	Mode	77.28	67.66	74.03	63.31
	Median	77.85	75.01	77.14	75.10
	IQR	(55.83,101.60)	(53.68,99.74)	(56.19,100.10)	(54.61,99.18)
E2	Mode	76.39	88.41	84.95	89.53
	Median	95.70	92.34	93.37	94.12
	IQR	(70.32,122.50)	(66.56,121.70)	(69.14,121.20)	(70.51,122.40)
E3	Mode	97.78	99.63	101.85	107.09
	Median	109.60	109.60	109.00	110.70
	IQR	(84.66,137.00)	(87.19,140.00)	(85.17,138.70)	(86.78, 138.30)
E4	Mode	90.61	93.05	80.92	84.32
	Median	99.47	98.45	98.00	99.47
	IQR	(73.76,128.10)	(74.62,126.00)	(73.24,129.20)	(75.21,127.50)
E5	Mode	107.76	115.49	120.64	122.43
	Median	122.90	119.70	120.60	122.80
	IQR	(95.41,156.80)	(92.87,150.00)	(91.18,151.40)	(93.79,151.70)
E6	Mode	41.27	48.45	45.97	46.95
	Median	48.96	51.31	49.01	49.39
	IQR	(33.79,66.66)	(35.72,67.64)	(34.50,66.67)	(34.07,68.40)

Table 3.8: Simulation: summary of the log-Bayes factor for the cluster model with covariates (MWC) and without covariates (MWOC)

Ind	ρ	ℓ	log-Bayes factor: WC				log-Bayes factor: WOC			
			mode	Q_1	Q_2	Q_3	mode	Q_1	Q_2	Q_3
1.0	.01	27	-1.87	-1.98	-1.95	-1.82	-1.69	-2.04	-1.92	-1.62
1.0	.10	27	-1.31	-1.87	-1.68	-1.32	-0.88	-1.75	-1.50	-0.83
1.0	.30	27	-0.22 [†]	-1.67	-1.42	-0.62	3.53 [†]	-1.28	-0.54	1.71
1.2	.01	27	-0.97	-1.66	-1.20	-0.82	-1.79	-2.05	-1.92	-1.69
1.2	.10	27	0.20 [†]	-1.72	-1.06	0.18	-0.62 [†]	-1.73	-1.44	-0.76
1.2	.30	27	0.17 [†]	-1.50	-1.16	-0.13	4.60 [†]	-1.11	-0.20	3.16
1.4	.01	27	4.20 [†]	0.95	3.54	6.43	0.35 [†]	-1.35	-0.57	0.94
1.4	.10	27	5.12	0.23	2.98	7.86	2.73	-1.16	0.22	3.56
1.4	.30	27	8.95	-1.55	-0.10	6.41	13.12	-0.49	2.80	18.57

NOTE: Ind is the degree of dependence, ρ is the intracluster correlation, ℓ is the number of cluster and [†] indicates that the corresponding values are different. The sample size is 837.

Table 3.9: Simulation: summary of the log-Bayes factor for the cluster model with covariates (MWC) and without covariates (MWOC)

Ind	ρ	ℓ	log-Bayes factor: WC				log-Bayes factor: WOC			
			mode	Q_1	Q_2	Q_3	mode	Q_1	Q_2	Q_3
1.0	.01	54	-2.06	-2.25	-2.17	-1.94	-1.71	-2.23	-1.99	-1.53
1.0	.10	54	-1.51	-2.15	-2.02	-1.76	-0.94	-1.96	-1.64	-0.86
1.0	.30	54	1.00	-1.70	-1.27	2.08	2.86	-1.45	-0.68	1.38
1.2	.01	54	0.37 [†]	-1.44	-0.27	1.73	-1.92 [†]	-2.29	-2.12	-1.74
1.2	.10	54	1.26 [†]	-1.51	-1.02	0.51	-0.71 [†]	-1.91	-1.48	-0.55
1.2	.30	54	4.41	-1.75	-1.50	-0.50	9.25	-1.10	1.08	8.65
1.4	.01	54	13.41	9.23	12.37	18.57	3.08	-0.30	1.18	5.79
1.4	.10	54	12.34	2.34	9.75	17.80	7.83	0.04	3.62	11.13
1.4	.30	54	16.18	-1.54	-0.38	31.04	25.20	1.73	14.00	37.38

NOTE: Ind is the degree of dependence, ρ is the intracluster correlation, ℓ is the number of cluster and [†] indicates that the corresponding values are different. The sample size is 1674.

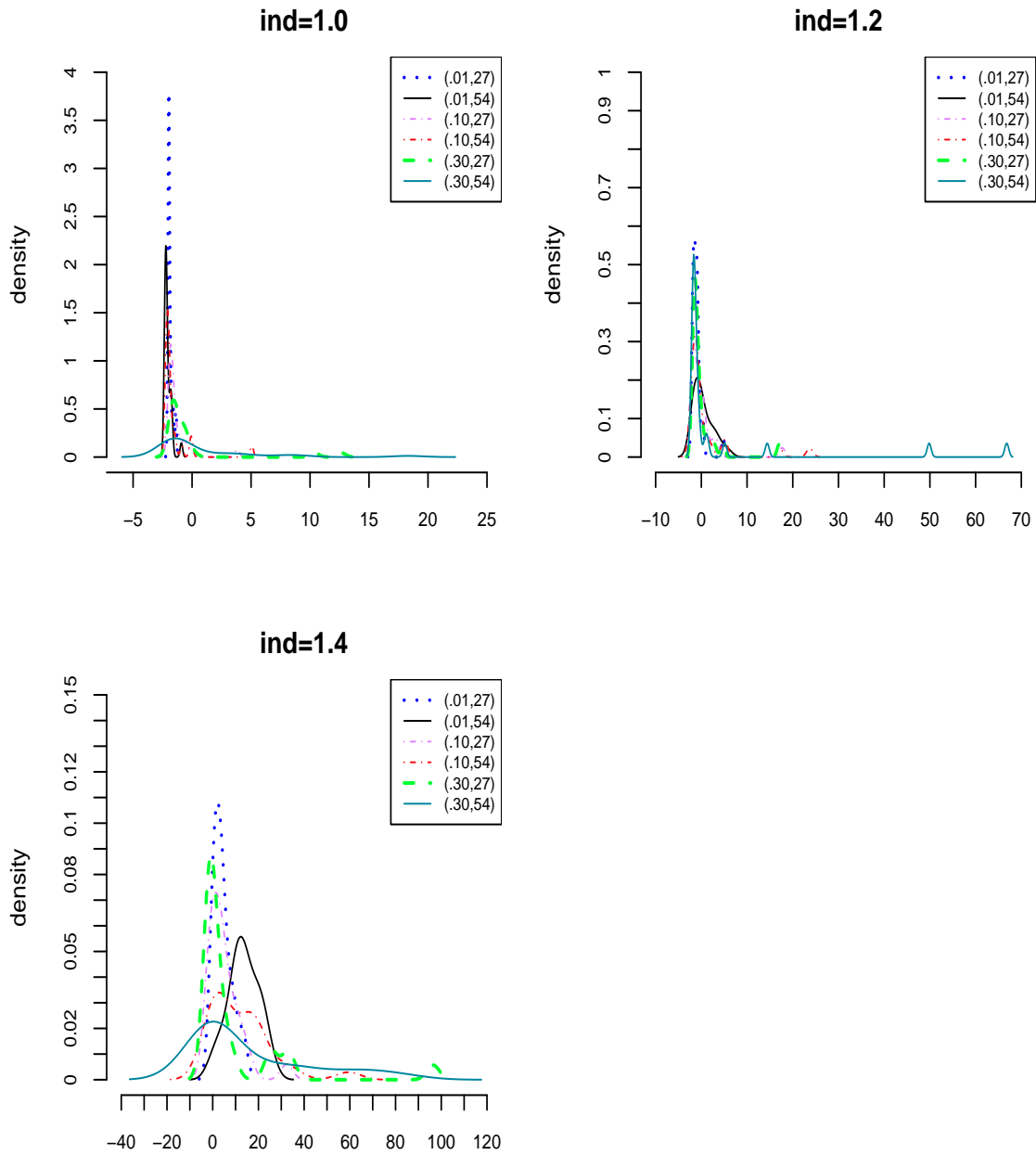


Figure 3.1: Plot of the empirical densities of the log-Bayes factors for the simulation with covariates when $\text{ind}=1.0$, $\text{ind}=1.2$ and $\text{ind}=1.4$. In the legend on the top right side, the first and second values of the pair represent the intraclass correlation (ρ) and the cluster size (ℓ) respectively.

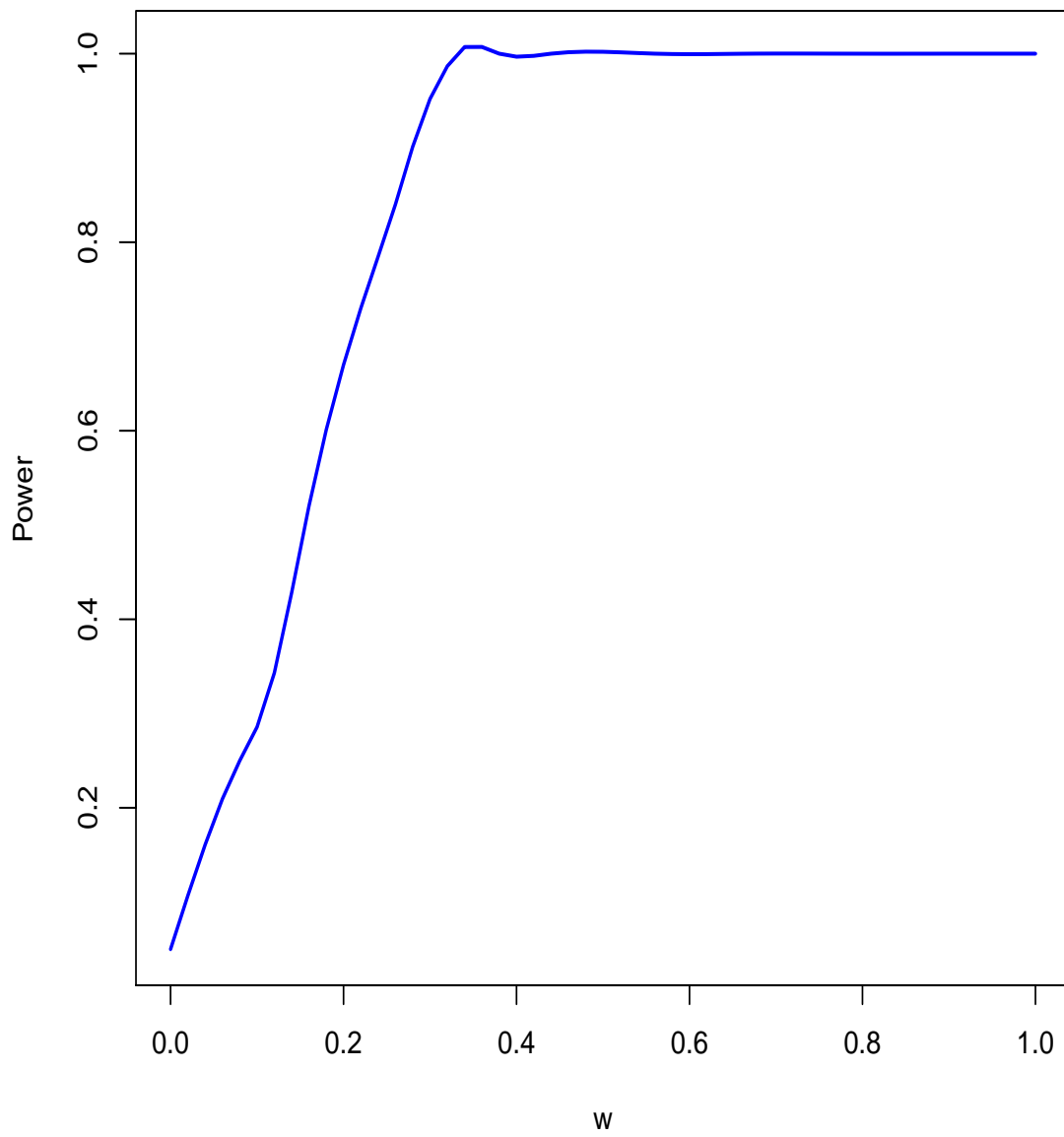


Figure 3.2: Plot of the estimated power function of the test

Chapter 4

Concluding Remarks

In Chapter 4, we summarize our methodological contributions and we discuss some future problems. We have developed a test of independence in two-way categorical tables for two-stage cluster sampling. We have applied our methods to two TIMSS data sets.

4.1 Contribution in Methodology

We have proposed methods to study independence in a two-way contingency table which has been obtained from two-stage cluster sampling design with simple random sampling at both stages. In doing so, we have studied an association (not directional) between two categorical variables when (a) there are no covariates and (b) there are covariates at unit level and/or cluster level.

In Chapter 2, for the test of independence without covariates, methodology is developed to overcome the limitations of Rao-Scott correction. The Rao-Scott methods were developed to correct for design effects such as cluster effects by correcting the standard Pearson's chi-squared (X^2) and the likelihood ratio (G^2) statistics. They are "large sample" methods and work well when there are large cell counts. However, they are less successful when the cell counts are small. We have used a hierarchical Bayesian model to convert the observed cluster samples to an equivalent simple random sample. This provides surrogate samples which can be used to derive the distribution of the Bayes factor to make an inference about independence. We use a sampling-based method to fit the model under which we draw a large number of samples from the approximate posterior density and subsample them using SIR algorithm. Although our method is a sampling based method, it is at least as fast as the Rao-

Scott methods. We demonstrate the utility of our procedure using an example from TIMSS 1995 to study the association between student's mathematics score and the community the student come from, and the student's science score and the community the student come from. We have also provided a simulation study that establishes our methodology as a viable alternative to the Rao-Scott approximations for relatively small two-stage cluster samples. Relative to standard methods, our approach provides additional insight by displaying the distribution of the Bayes factor rather than simply relying on a single summary measure.

In Chapter 3, for the test of independence with covariates, we have developed a model in which we incorporate the covariates and accommodated the clustering effect. We have used an idea of surrogate sampling similar to the one applied in Chapter 2. However, in this case, the cluster sample with covariates is first converted to a cluster sample without covariates. We then have converted this cluster sample to an equivalent simple random sample using the hierarchical Bayesian model, which is used to compute the Bayes factor to make inference about the test of independence. The second part of the procedure is similar to what was done in Chapter 2. However, we have used a new methodology here rather than adopting the procedure from Chapter 2 because the procedure presented in Chapter 2 is expensive for this new problem. We have used a Gibbs sampling method to fit the model. We have demonstrated the utility of our procedure using examples and also provide a simulation study. We have fitted both models (i) with covariates and (ii) without covariates. The results show that if there is a strong association between two categorical variables, there is no difference in an inference for the test of independence between two models. However, there is a noticeable difference in the corresponding inferences between the two models when there are borderline cases (i.e., situations where there is marginal significance).

Although we developed methods for the test of independence in two-way categorical tables for two-stage cluster sampling with simple random sampling at both stages, the methods are more general and can also work for two-stage cluster sampling with proportional to population size (pps) sampling at both stages, in which case we have a self-weighting sample of units.

Finally, we note that, although we applied surrogate sampling to the test of independence

using some specific examples for educational data, there are many other applications of surrogate sampling. One example is data masking, which is the process of obscuring (masking) sensitive data by replacing it with realistic but not real data in order to reduce the exposure of sensitive information. In many government agencies and research organizations, due to confidentiality issues, real data is barely published. Our surrogate sampling can be used as a data masking procedure so that the surrogate sample can be made available to secondary data analysts.

4.2 Interpretations of Surrogate Sampling Table and discussion

It is interesting to compare the surrogate data and the observed data. For illustration we have chosen the model and the examples (TIMSS 1995 data) from Chapter 2 because these examples have substantial clustering effects. In Chapter 2, for the Bayesian test of independence, we have converted the total cluster table into a large number of equivalent simple random samples which are the surrogates of the original data.

To simplify the discussion, we average the surrogate tables to obtain a single table which we call TSUR. We call the total table for the observed data TOBS. We have also obtained two more total tables which use the effective sample sizes, one for the observed data and another for the surrogate data. The first table, which we call ETOBS, is obtained by dividing the cells count in TOBS by the Bayesian design effects (BDEFs), presented in Table 2.2. To get the second table, which we call ETSUR, we divide the cell counts in each of the surrogate tables by the same BDEFs, and we average all the tables to obtain a single table. We present all these tables for examples E1-E8 in Table 4.1.

Our main interest is to compare TOBS and TSUR. In Table 4.1 as we expect, there are some differences in the cell counts between these tables. For example, in E1, TOBS has 57, 5, 83, 63 respectively in $(1, 2)$, $(1, 3)$, $(2, 1)$, $(3, 1)$ and TSUR has 28, 34, 43, 83. Similarly, in E4, TOBS has 17, 157, 134, 294 in $(1, 2)$, $(2, 1)$, $(2, 2)$, $(3, 2)$ and TSUR has 46, 252, 95, 190. Thus, the surrogate samples are different from the observed tables. This is due to large intracluster correlations in these examples.

We performed the test of independence on each of the tables (TOBS, ETOBS, TSUR and ETSUR) for all examples E1-E8. We compute the Bayes factors and the p-values for the chi-squared and likelihood ratio test; the results are presented in Table 4.2. Note here that the test based on the total observed table is not the right test because it is not adjusted with the design effects. However, the tests based on the other tables are expected to be correct; the counts in TSUR may be too large though. There are some agreement and disagreement among the χ^2 test, the G^2 test and the Bayesian test. For example in E3 and E6, they mostly agree with each other inferring that there is evidence of independence between math score and communities in E3 and between science score and communities in E6. However, they do not all agree in rest of the examples. But the tests based on ETOBS and ETSUR do mostly agree in all examples.

Dividing the total table TOBS by the design effects to get ETOBS is similar to what Rao and Scott (1981) did. We believe that TSUR should not have the same sample size as the observed data because the observed data are correlated owing to the clustering effect but the simple random simple data are not correlated. So there may be excessive information in TSUR. We also divide the cells by the same design effects, and currently, it is not clear what is the best way to proceed. We contemplate working in this problem in future.

Table 4.1: Comparison of the observed total table and the surrogate total table by example

Example	Table	(1, 1)	(1, 2)	(1, 3)	(2, 1)	(2, 2)	(2, 3)	(3, 1)	(3, 2)	(3, 3)	Total
E1	TOBS	44	57	5	83	71	5	63	136	5	469
	ETOB	5	8	0	10	10	0	10	20	0	63
	TSUR	45	28	34	43	78	8	83	142	8	469
	ETSUR	6	4	2	5	11	0	13	21	0	62
E2	TOBS	49	74	1	107	151	13	93	164	11	663
	ETOB	9	14	0	20	29	2	17	30	1	122
	TSUR	36	81	15	84	166	26	112	109	34	663
	ETSUR	6	15	0	15	32	4	21	20	5	118
E3	TOBS	44	47	8	54	44	3	56	167	15	438
	ETOB	9	9	1	11	8	0	10	34	2	84
	TSUR	45	55	13	27	49	4	77	130	38	438
	ETSUR	9	11	1	5	8	0	14	26	6	80
E4	TOBS	25	17	0	157	134	13	205	294	12	857
	ETOB	4	2	0	30	26	2	38	57	1	160
	TSUR	38	46	4	252	95	28	181	190	23	857
	ETSUR	7	8	0	48	18	4	33	36	3	157
E5	TOBS	63	38	5	105	47	7	70	124	10	469
	ETOB	9	5	0	13	7	0	10	18	0	62
	TSUR	58	24	5	115	45	17	54	137	14	469
	ETSUR	8	3	0	15	7	1	8	20	1	63
E6	TOBS	61	56	7	117	141	13	117	145	6	663
	ETOB	12	12	0	23	27	2	23	27	0	126
	TSUR	45	37	9	170	126	34	136	90	16	663
	ETSUR	9	8	1	33	24	5	27	16	1	124
E7	TOBS	53	44	2	67	30	4	95	133	10	438
	ETOB	11	8	0	13	5	0	16	25	1	79
	TSUR	52	25	9	52	29	20	102	133	16	438
	ETSUR	11	4	0	10	5	2	17	25	2	76
E8	TOBS	34	7	1	181	112	11	226	272	13	857
	ETOB	6	0	0	34	21	1	41	52	2	157
	TSUR	66	21	2	183	77	19	239	229	21	857
	ETSUR	12	2	0	34	15	2	44	43	3	155

Table 4.2: Comparison of inference from the observed total table and the surrogate total table by example

Example	Table	p-values		log-BF
		χ^2	G^2	
E1	TOBS	0.00078	0.00078	1.76
	ETOBS*	0.78126	0.78681	-2.83
	TSUR	0.00000	0.00000	28.01
	ETSUR	0.29211	0.32900	-1.33
E2	TOBS	0.24716	0.15015	-5.54
	ETOBS	0.96669	0.96454	-4.87
	TSUR	0.00035	0.00033	2.94
	ETSUR	0.25675	0.25096	-1.95
E3	TOBS	0.00000	0.00000	10.09
	ETOBS	0.06817	0.06547	0.09
	TSUR	0.07659	0.05312	-2.40
	ETSUR	0.76671	0.75176	-3.06
E4	TOBS	0.00140	0.00145	-1.34
	ETOBS	0.07041	0.14045	-3.16
	TSUR	0.00000	0.00000	16.78
	ETSUR	0.04342	0.03962	-0.66
E5	TOBS	0.00000	0.00000	13.65
	ETOBS*	0.24715	0.24273	-0.95
	TSUR	0.00000	0.00000	34.61
	ETSUR	0.02922	0.02612	1.67
E6	TOBS	0.23971	0.22037	-5.45
	ETOBS	0.97307	0.97070	-4.96
	TSUR	0.51144	0.49650	-6.17
	ETSUR	0.88711	0.88628	-4.35
E7	TOBS	0.00020	0.00015	3.44
	ETOBS	0.17346	0.16263	-1.05
	TSUR	0.00000	0.00000	9.20
	ETSUR	0.12072	0.11850	-0.33
E8	TOBS	0.00000	0.00000	8.44
	ETOBS	0.02906	0.04128	-1.80
	TSUR	0.00000	0.00000	11.03
	ETSUR	0.05706	0.04405	-1.27

Note: In Examples E1 and E2, * indicates that ETOBS has all zeros in its last column.

4.3 Future Work

The following problems can be solved within our framework.

4.3.1 Stratified Two-stage Cluster Sampling

It is easy to accommodate stratification in our framework because this is simply an additional step in our two-stage cluster sampling procedure. We just need to index all quantities with s (for stratum). For instance, if we consider our model for a two-stage cluster sampling design without covariates in (2.2), (2.3) and (2.4), we can write the model of stratified two-stage cluster design as

$$\eta_{hi} \sim \text{Multinomial}(n_{hi}, \underline{a}_{hi}), \quad (4.1)$$

where $\eta_{hi} = (n_{hi1}, \dots, n_{hi\ell})$, $n_{hi} = \sum_{s=1}^S n_{his}$ and $a_{his} = \alpha_{his}\pi_s$, $h = 1, \dots, H$, $i = 1, \dots, \ell$, $s = 1, \dots, S$. Here, the indices h , i and s stand for stratum, cluster and cell of the table respectively. Note here that $\alpha_{his}\pi_s$ is the probability that a unit has the s^{th} characteristic within the i^{th} cluster of stratum h of the super population, and π_s , $s = 1, \dots, S$, are the probabilities corresponding to a homogeneous superpopulation (i.e., there are no strata and clusters). We want the test of independence based on π_s . In (4.1) we have the constraints $a_{his} = \alpha_{his}\pi_s$, $\sum_s \alpha_{his}\pi_s = 1$, $\sum_s \pi_s = 1$, $\alpha_{his}\pi_s > 0$ and $\pi_s > 0$. Here, the α_{his} are used to adjust for the clustering. A priori we take

$$\alpha_{his} \stackrel{iid}{\sim} \text{Gamma}(\tau_{hs}, \tau_{hs}\nu_h), \quad (4.2)$$

$$\underline{\pi} \sim \text{Dirichlet}(\underline{1}), \quad (4.3)$$

and

$$p(\nu_h) \propto 1/\nu_h, \quad h = 1, \dots, H \quad \text{independent}, \quad (4.4)$$

where $\underline{\pi} = (\pi_1, \dots, \pi_S)$ and τ_{hs} , $h = 1, \dots, H$, $s = 1, \dots, S$ are to be specified. We can apply the same idea for the cluster model with covariates.

4.3.2 Introduction of Survey Weights

Let $\{w_{ij}, i = 1, \dots, \ell, j = 1, \dots, n_i\}$ denote the survey weights in a two-stage cluster sampling design. Each w_{ij} is the number (including itself) that each sampled unit represents in

the population. Thus, these w_{ij} sum to the total population size. There are many controversies of how survey weights should be handled (e.g., Gelman, 2007). We will mention one possible method that we will consider using.

We plan to incorporate the survey weights partially using the technique of Pfeffermann, Skinner, Holmes, Goldstein and Rashbash (1998) who devised the scheme for applying sampling weights to the likelihood function. In fact, Pfeffermann et al. (1998) applied sampling weights to multilevel samples by defining a pseudo-likelihood. The resulting multilevel pseudo-likelihood is maximized to yield maximum likelihood estimates of the model parameters (see Pfeffermann, et al, 1998). Under regularity conditions, pseudo-likelihood estimators are consistent and asymptotically normal (Arnold and Strauss, 1991). However, pseudo-likelihood is an approximate likelihood and, of course, a Bayesian will use proper likelihood.

Under normal distribution the pseudo-likelihood approach is the same as the correct likelihood approach. Otherwise these two are different as discussed below using two examples.

Example 1: Let $Y \sim N(\mu, \sigma^2)$ and w be the sampling weight associated with the unit. The density function of the random variable is

$$p(y) = \frac{1}{\sqrt{(2\pi\sigma^2)}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}.$$

For a single unit the pseudo-likelihood after applying the sampling weight is

$$L_1(y) = [p(y)]^w = \left[\frac{1}{\sqrt{(2\pi\sigma^2)}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} \right]^w, \quad (4.5)$$

and the exact likelihood function is

$$L_2(y) = \frac{[p(y)]^w}{\int [p(y)]^w dy} = \frac{\left[\frac{1}{\sqrt{(2\pi\sigma^2)}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} \right]^w}{\int \left[\frac{1}{\sqrt{(2\pi\sigma^2)}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} \right]^w dy}. \quad (4.6)$$

It is easy to show that the estimation of the parameters using the likelihood in (4.5) is equivalent to the one in (4.6).

Example 2: Let us consider a logistic function

$$p(y) = \frac{e^{x'\beta y}}{1 + e^{x'\beta y}},$$

where $y = 0$ or 1 . For a single unit, the pseudo-likelihood after applying the sampling weight is

$$L_1(y) = [p(y)]^w = \left[\frac{e^{x' \beta y}}{1 + e^{x' \beta}} \right]^w, \quad (4.7)$$

and the exact likelihood function is

$$L_2(y) = \frac{[p(y)]^w}{\sum_y [p(y)]^w} = \frac{\left[\frac{e^{x' \beta y}}{1 + e^{x' \beta}} \right]^w}{\sum_y \left[\frac{e^{x' \beta y}}{1 + e^{x' \beta}} \right]^w} = \frac{\frac{e^{x' \beta w}}{[1 + e^{x' \beta}]^w}}{\frac{e^{x' \beta w}}{[1 + e^{x' \beta}]^w} + \frac{1}{[1 + e^{x' \beta}]^w}} = \frac{e^{x' \beta w}}{1 + e^{x' \beta w}}. \quad (4.8)$$

Clearly, the estimation of the parameters using the likelihood in (4.7) is different from the one in (4.8). This can be easily seen by showing that normal equations and Hessian matrices for the estimation of the parameters are different. We will extend the idea explained in Example 2 for our cluster model as described briefly below.

Let us consider the model from Chapter 2:

$$n_i \mid a_i \stackrel{ind}{\sim} \text{Multinomial}(n_i, a_i),$$

where $a_{is} = \alpha_{is} \pi_s$, $s = 1, \dots, S$. This is equivalent to

$$I_{ij} \mid a_i \stackrel{iid}{\sim} \text{Multinomial}(1, a_i),$$

where I_{ij} refers to the cell in which j^{th} individual falls in the i^{th} cluster with $\sum_{s=1}^S I_{ijs} = 1$ and $\sum_{j=1}^{n_i} \sum_{s=1}^S I_{ijs} = n_i$. Including the survey weights $\{w_{ij}, i = 1, \dots, \ell, j = 1, \dots, n_i\}$

$$P(I_{ij} \mid a_i) = \frac{[\prod_{s=1}^S (\alpha_{is} \pi_s)^{I_{ijs}}]^{w_{ij}}}{\sum_{\{I_{ijs}\}} [\prod_{s=1}^S (\alpha_{is} \pi_s)^{I_{ijs}}]^{w_{ij}}} = \frac{[\prod_{s=1}^S (\alpha_{is} \pi_s)^{I_{ijs}}]^{w_{ij}}}{\sum_{s=1}^S (\alpha_{is} \pi_s)^{w_{ij}}}.$$

Then,

$$P(I \mid a) = \prod_{i=1}^{\ell} \prod_{j=1}^{n_i} \frac{[\prod_{s=1}^S (\alpha_{is} \pi_s)^{I_{ijs}}]^{w_{ij}}}{\sum_{s=1}^S (\alpha_{is} \pi_s)^{w_{ij}}}.$$

The rest of the model is the same as in Chapter 2. Also, the computation can be done in a very similar manner.

4.3.3 Sampling Zero Problem

Here, we discuss our plan to deal with a problem when there are sampling zeros in cluster tables by using the Brier (1980) model. We will extend this idea to our cluster model later.

When a cluster sampling is performed, there will be many cells with zero counts. The total counts in each cluster are generally assumed fixed and known, so when a parametric model is fitted to the data, the fitted values corresponding to the zero counts get much larger than zeros and the positive counts get smaller. This causes the model studied by Brier (1980) to fit poorly. The Brier (1980) model is

$$\underline{n}_i | \underline{\pi}_i \stackrel{ind}{\sim} \text{Mult}(n_i, \underline{\pi}_i), \quad \underline{\pi}_i | \underline{\mu}, \tau \stackrel{iid}{\sim} \text{Dirichlet}(\underline{\mu}\tau), \quad i = 1, \dots, \ell,$$

where $\underline{n}_i = (n_{i1}, \dots, n_{iS})$ and $\underline{\pi}_i = (\pi_{i1}, \dots, \pi_{iS})$. Let $\mathcal{C}_i = \{s : n_{is} > 0\}$ and $\bar{\mathcal{C}}_i = \{s : n_{is} = 0\}$. Let

$$z_{is} = \begin{cases} 0, & -1, \quad s \in \mathcal{C}_i \\ 0, & 1, \quad s \notin \mathcal{C}_i, \end{cases}$$

where $s = 1, \dots, S$. That is, we plan to remove one observation or leave unchanged the positive cell and for a zero cell we either add an observation or leave it unaltered. We require $\sum_{s=1}^S (n_{is} + z_{is}) = n_i$, i.e., $\sum_{s=1}^S n_{is} + \sum_{s=1}^S z_{is} = n_i$ and $\sum_{s=1}^S n_{is} = n_i$, so that

$$P(\underline{n}_i, \underline{z}_i | \underline{\pi}_i) = n_i! \left\{ \prod_{s \in \mathcal{C}_i} \frac{\pi_{is}^{n_{is} + z_{is}}}{(n_{is} + z_{is})!} \times \prod_{s \notin \mathcal{C}_i} \frac{\pi_{is}^{z_{is}}}{z_{is}!} \right\}.$$

The likelihood function of the data is

$$P(\underline{n}, \underline{z} | \underline{\pi}) = \prod_{i=1}^{\ell} \left\{ n_i! \prod_{s \in \mathcal{C}_i} \frac{\pi_{is}^{n_{is} + z_{is}}}{(n_{is} + z_{is})!} \times \prod_{s \notin \mathcal{C}_i} \frac{\pi_{is}^{z_{is}}}{z_{is}!} \right\}, \quad (4.9)$$

where $\sum_{s=1}^S n_{is} = n_i$, $\sum_{s=1}^S z_{is} = 0$ and $n_{is} = 0$ for $s \notin \mathcal{C}_i$.

A priori we assume

$$\begin{aligned} \underline{\pi}_i | \underline{\mu}, \rho &\sim \text{Dirichlet} \left(\underline{\mu} \frac{1 - \rho}{\rho} \right), \\ \pi(\underline{\mu}, \rho) &= 1. \end{aligned} \quad (4.10)$$

Now, combining the likelihood function in (4.9) and priors in (4.10) via Bayes' theorem, we obtain the joint posterior density

$$\pi(\underline{z}, \underline{\pi}, \underline{\mu}, \rho | \underline{n}) \propto \prod_{i=1}^{\ell} \left\{ n_i! \prod_{s \in \mathcal{C}_i} \frac{\pi_{is}^{n_{is} + z_{is}}}{(n_{is} + z_{is})!} \times \prod_{s \notin \mathcal{C}_i} \frac{\pi_{is}^{z_{is}}}{z_{is}!} \times \frac{\prod_{s=1}^S \pi_{is}^{\mu_s (\frac{1-\rho}{\rho}) - 1}}{D(\underline{\mu} \frac{1-\rho}{\rho})} \right\}. \quad (4.11)$$

This method does not cover a single large table with many zeros because the probabilities

of the zero cells can not be estimated efficiently. A uniform prior is not reasonable for π in a large table with many zeros. A likelihood ratio test of independence in a single contingency table is given by Nandram, Bhatta and Bhadra (2012) with many sampling zeros under simple random sampling. We will apply this method to two-stage cluster sampling.

4.4 My Accomplishments

Besides my dissertation, I have also worked on various other problems and have submitted corresponding papers. The work in Chapter 2 is reported in Nandram, Bhatta, Bhadra and Sedransk (2012). Two of the papers have been published for the publication and the other two are still under review. I will briefly discuss these works here.

4.4.1 Mortality Curve Fitting

a. Switching Nonlinear Regression Model

Bhatta and Nandram (2013) considered fitting of age-specific mortality curve to English and Welsh (1988-1992) mortality data. We used the eight-parameter Heligman-Pollard (HP) empirical law to fit the mortality curve. It consists of three nonlinear curves: child mortality, mid-life mortality and adult mortality. The eight unknown parameters in the HP law are difficult to estimate because of a convergence problem during computation. In order to overcome this problem, we considered a novel idea to fit the three curves (nonlinear splines) separately, and then connect them smoothly at the two knots. To connect the curves smoothly, we express uncertainty about the knots because these curves do not have turning points.

b. Small Area

Wei, Nandram and Bhatta (2012) considered fitting of mortality curves to US mortality data for 1999-2001 summed across all ages in each race-gender domain (white males, black males, white females and black females) by state. We used the eight-parameter Heligman-Pollard (HP) empirical law to fit these curves. Because the data are studied in small domains, the death counts for some ages tend to be very low, and for some

small areas, few or zero deaths can be observed in a time period. This causes difficulties for fitting the HP law to accurately estimate real mortalities and model age-mortality curves in smoothing patterns for those areas. Our Bayesian method provides a solution to overcome these difficulties by using data from other states.

4.4.2 Selection Bias

Nandram, Bhatta, Bhadra and Shen (2012) have shown how to infer about a finite population proportion using data from a possibly biased sample. We have used the Bayesian nonignorable selection model to accommodate the selection mechanism. We have extended the work of Malec, Davis and Cao (1999) in a direction different from that of Nandram and Choi (2010). We illustrated our method using numerical examples obtained from NHIS 1995 data. The result shows that our nonignorable selection model appears to accommodate the selection mechanism reasonably well.

4.4.3 Sparse Two-Way Contingency Tables

Nandram, Bhatta and Bhadra (2012) considered a likelihood ratio test of independence for large two-way contingency tables having cells with small and/or zero counts. Specifically, we restricted attention to tables with many sampling (random) zeros which can become positive with larger sample sizes. We combined all the cells with sampling zeros to form a single positive cell. Then, assuming all cell counts are positive random variables, we modeled the counts using a truncated multinomial distribution, thereby providing a test of quasi-independence for two-way contingency tables. In fact, we have two truncated multinomial distributions; one of these is for the null hypothesis of independence and the other for the unrestricted parameter space.

Appendix A

Joint Posterior Density

Letting $S = rc$, the set of constraints is

$$T = \left\{ (\underline{\alpha}, \underline{\pi}, \nu) : \sum_{s=1}^S \alpha_{is} \pi_s = 1, \sum_{s=1}^S \pi_s = 1, \alpha_{is} > 0, i = 1, \dots, \ell, \pi_s > 0, s = 1, \dots, S, s\nu > \nu_o \right\}.$$

Letting $b = \sum_{s=1}^S \tau_s$, the joint prior density is

$$p(\underline{\alpha}, \underline{\pi}, \nu \mid \underline{\tau}) \propto \nu^{\ell b - 1} \prod_{i=1}^{\ell} \prod_{s=1}^S \alpha_{is}^{\tau_s - 1} e^{-\nu \tau_s \alpha_{is}}, \quad (\underline{\alpha}, \underline{\pi}, \nu) \in T. \quad (\text{A.1})$$

In (A.1) we want to accommodate the constraints, $\sum_{s=1}^S \alpha_{is} \pi_s = 1$, $i = 1, \dots, \ell$, and $\sum_{s=1}^S \pi_s = 1$. We have a convenient way of doing so.

We transform α_{iS} , $i = 1, \dots, \ell$, to ϕ_i and π_S to ϕ_0 , keeping all other random variables untransformed so that

$$\sum_{s=1}^S \alpha_{is} \pi_s = 1 + \phi_i, \quad i = 1, \dots, \ell \quad \text{and} \quad \sum_{s=1}^S \pi_s = 1 + \phi_0.$$

Our idea is to remove π_S and α_{iS} , $i = 1, \dots, \ell$, when ϕ_i , $i = 0, 1, \dots, \ell$, are set to zero. Then, $\pi_S = 1 + \phi_0 - \sum_{s=1}^{S-1} \pi_s$ and $\alpha_{iS} = \frac{1 + \phi_i - \sum_{s=1}^{S-1} \alpha_{is} \pi_s}{1 + \phi_0 - \sum_{s=1}^{S-1} \pi_s}$, $i = 1, \dots, \ell$. Note that π_S and α_{iS} are all kept in $(0, 1)$.

The Jacobian of the transformation is $\left(\left| 1 + \phi_0 - \sum_{s=1}^{S-1} \pi_s \right| \right)^{-\ell}$ and the joint prior density is

$$p(\underline{\alpha}_{(S)}, \underline{\pi}_{(S)}, \underline{\phi}, \nu) \propto \nu^{\ell b - 1} \prod_{i=1}^{\ell} \left[\prod_{s=1}^{S-1} \alpha_{is}^{\tau_s - 1} e^{-\nu \tau_s \alpha_{is}} \right. \\ \left. \times \left(\frac{1 + \phi_i - \sum_{s=1}^{S-1} \alpha_{is} \pi_s}{1 + \phi_0 - \sum_{s=1}^{S-1} \pi_s} \right)^{\tau_S - 1} e^{-\nu \tau_S \left(\frac{1 + \phi_i - \sum_{s=1}^{S-1} \alpha_{is} \pi_s}{1 + \phi_0 - \sum_{s=1}^{S-1} \pi_s} \right)} \left(\left| 1 + \phi_0 - \sum_{s=1}^{S-1} \pi_s \right| \right)^{-1} \right],$$

$$0 < \sum_{s=1}^{S-1} \alpha_{is} \pi_s < 1, \quad i = 1, \dots, \ell, \quad 0 < \sum_{s=1}^{S-1} \pi_s < 1, \quad \alpha_{is} \pi_s > 0, \quad \pi_s > 0, \quad \nu \geq \pi^*.$$

Then, letting

$$\begin{aligned} \tilde{T} = & \left\{ (\alpha_{(S)}, \pi_{(S)}, \nu) : 0 < \sum_{s=1}^{S-1} \alpha_{is} \pi_s < 1, i = 1, \dots, \ell, 0 < \sum_{s=1}^{S-1} \pi_s < 1, \alpha_{is} \pi_s > 0, \pi_s > 0, \right. \\ & \left. s = 1, \dots, S-1, \nu > \nu_o \right\}, \\ p(\alpha_{(S)}, \pi_{(S)}, \nu \mid \phi = 0) \propto & \nu^{\ell b-1} \prod_{i=1}^{\ell} \left[\prod_{s=1}^{S-1} \alpha_{is}^{\tau_s-1} e^{-\nu \tau_s \alpha_{is}} \right. \\ & \left. \times \left(\frac{1 - \sum_{s=1}^{S-1} \alpha_{is} \pi_s}{1 - \sum_{s=1}^{S-1} \pi_s} \right)^{\tau_s-1} e^{-\nu \tau_s \left(\frac{1 - \sum_{s=1}^{S-1} \alpha_{is} \pi_s}{1 - \sum_{s=1}^{S-1} \pi_s} \right)} \left(1 - \sum_{s=1}^{S-1} \pi_s \right)^{-1} \right], (\alpha_{(S)}, \pi_{(S)}, \nu) \in \tilde{T}. \end{aligned} \quad (\text{A.2})$$

Henceforth, for convenience, we will denote this prior distribution by $p(\alpha_{(S)}, \pi_{(S)}, \nu)$ which, we note, is improper.

Now, the conditional distribution of η given $(\alpha_{(S)}, \pi_{(S)}, \nu) \in \tilde{T}$ is

$$p(\eta \mid \alpha_{(S)}, \pi_{(S)}, \nu) = \prod_{i=1}^{\ell} \left[n_i! \left(\prod_{s=1}^{S-1} (\alpha_{is} \pi_s)^{n_{is}} / n_{is}! \right) \left(1 - \sum_{s=1}^{S-1} \alpha_{is} \pi_s \right)^{n_{iS}} / n_{iS}! \right] \quad (\text{A.3})$$

$$n_{is} \geq 0, \sum_{s=1}^S n_{is} = n_i, i = 1, \dots, \ell.$$

Then, using Bayes' theorem, the joint posterior density is

$$\begin{aligned} p(\alpha_{(S)}, \pi_{(S)}, \nu \mid \eta) \propto & \prod_{i=1}^{\ell} \left[n_i! \left(\prod_{s=1}^{S-1} (\alpha_{is} \pi_s)^{n_{is}} / n_{is}! \right) \left(1 - \sum_{s=1}^{S-1} \alpha_{is} \pi_s \right)^{n_{iS}} / n_{iS}! \right] \\ & \times \nu^{\ell b-1} \prod_{i=1}^{\ell} \left[\prod_{s=1}^{S-1} \alpha_{is}^{\tau_s-1} e^{-\nu \tau_s \alpha_{is}} \right. \\ & \left. \times \left(\frac{1 - \sum_{s=1}^{S-1} \alpha_{is} \pi_s}{1 - \sum_{s=1}^{S-1} \pi_s} \right)^{\tau_s-1} e^{-\nu \tau_s \left(\frac{1 - \sum_{s=1}^{S-1} \alpha_{is} \pi_s}{1 - \sum_{s=1}^{S-1} \pi_s} \right)} \left(1 - \sum_{s=1}^{S-1} \pi_s \right)^{-1} \right], (\alpha_{(S)}, \pi_{(S)}, \nu) \in \tilde{T}. \end{aligned} \quad (\text{A.4})$$

Note that in (A.4) $\alpha_{iS} = (1 - \sum_{s=1}^{S-1} \alpha_{is} \pi_s) / (1 - \sum_{s=1}^{S-1} \pi_s)$ and $\pi_S = 1 - \sum_{s=1}^{S-1} \pi_s$.

Appendix B

A Property of the Gamma Distribution

Let $d_1, \dots, d_n \stackrel{iid}{\sim} \text{Gamma}(e, ef)$. Let $A = \sum_{i=1}^n d_i/n$ and $G = (\prod_{i=1}^n d_i)^{1/n}$ denote respectively the arithmetic and the geometric mean of the d_i .

Lemma

The maximum likelihood estimator (MLE) of f is $\hat{f} = A^{-1}$ which is the unique solution of

$$\ln(\hat{f}) - \psi(\hat{f}) = \ln(A/G), \quad (\text{B.1})$$

where $\psi(\cdot)$ is the digamma function.

Proof of Lemma

The log-likelihood function is

$$\Delta(e, f) = n\{e \ln(f) + e \ln(e) + (e - 1) \ln(G) - efA - \ln(\Gamma(e))\}.$$

Differentiating, we have,

$$\frac{\partial \Delta(e, f)}{\partial f} = ne \left(\frac{1}{f} - A \right) \quad \text{and} \quad \frac{\partial^2 \Delta(e, f)}{\partial f^2} = -\frac{ne}{f^2}. \quad (\text{B.2})$$

Using (B.2) it follows that the MLE of f is unique and is given by $\hat{f} = A^{-1}$.

Thus, the profile log-likelihood is

$$\Delta(e, \hat{f}) = n\{e \ln(\hat{f}) + e \ln(e) + (e - 1) \ln(G) - e - \ln(\Gamma(e))\}.$$

Differentiating, we have,

$$\frac{\partial \Delta(e, \hat{f})}{\partial e} = n \{ \ln(e) - \psi(e) + \ln(G/A) \} \quad \text{and} \quad \frac{\partial^2 \Delta(e, \hat{f})}{\partial e^2} = \frac{1}{e} - \psi'(e), \quad (\text{B.3})$$

where $\psi'(\cdot)$ is the trigamma function.

Then, because $e\psi'(e) > 1$ for all positive real numbers e (Abramowitz and Stegun 1969, Ch. 6), it follows from (B.3) that the MLE of e is the unique solution of (B.1).

Appendix C

Mode of a Kernel Density Estimator

Let $x_1, \dots, x_n \stackrel{iid}{\sim} f(x)$, where $f(x)$ is an unknown density function. We need the mode of this density function based on a large sample of size n . We use the Parzen-Rosenblatt kernel density estimator with a standard normal kernel and optimal window width (Silverman 1986), where

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n \phi\left(\frac{x-x_i}{h}\right), \quad -\infty < x < \infty, \quad (\text{C.1})$$

and h is the optimal window width.

Using differentiation,

$$\hat{f}'(x) = -\frac{1}{nh^3} \sum_{i=1}^n (x-x_i) \phi\left(\frac{x-x_i}{h}\right)$$

and

$$\hat{f}''(x) = -\frac{1}{nh^3} \sum_{i=1}^n \left\{1 - \left(\frac{x-x_i}{h}\right)^2\right\} \phi\left(\frac{x-x_i}{h}\right).$$

A necessary condition for a mode x^* is that $\hat{f}'(x^*) = 0$, which gives

$$x^* = \sum_{i=1}^n w\{(x^* - x_i)\} x_i, \quad (\text{C.2})$$

where $w\{(x^* - x_i)\} = \phi\left(\frac{x^* - x_i}{h}\right) \left\{ \sum_{i=1}^n \phi\left(\frac{x^* - x_i}{h}\right) \right\}^{-1}$, $i = 1, \dots, n$ (i.e., x^* is a weighted average).

We use a simple iterative procedure to solve (C.2). Starting with the sample mean on the right side of (C.2), we update x^* and iterate the procedure. This procedure is very fast even though it can take a large number of iterations for convergence. We need to check that $\hat{f}''(x^*) < 0$. This is approximately true because $\left\{1 - \left(\frac{x-x_i}{h}\right)^2\right\} \approx \exp\left\{-\left(\frac{x-x_i}{h}\right)^2\right\}$ which is positive. In fact, it is easy to show that $\hat{f}''(x^*) \geq -h^{-1}$; so it can be negative.

Alternatively, the global mode can be found by drawing samples from (C.1) and then finding the maximum of the values of $\hat{f}(x)$ over these samples; this procedure is easy and fast.

We have performed both procedures and they give virtually the same answer; but the latter procedure is expected always to work (Robert and Casella 1999, Ch. 5) for more complex optimization procedures.

Appendix D

Joint Posterior Density: A Simplification

Here, we provide the algebra to simplify a term from joint posterior density (3.7). We show that

$$\begin{aligned}
& \sum_{i=1}^{\ell} (\nu_i - \gamma' z_i)^2 + (\gamma - \gamma_0)' \Delta_0^{-1} (\gamma - \gamma_0) \\
&= \gamma_0' \frac{(z'z)}{\kappa+1} \gamma_0 + \nu' \left[I - \frac{\kappa}{\kappa+1} z(z'z)^{-1} z' \right] \nu - \frac{2}{\kappa+1} \gamma_0' z' \nu \\
& \quad + \left(\gamma - \frac{\kappa \hat{\gamma} + \gamma_0}{\kappa+1} \right)' \left(\frac{\kappa+1}{\kappa} \right) \hat{\Delta}^{-1} \left(\gamma - \frac{\kappa \hat{\gamma} + \gamma_0}{\kappa+1} \right).
\end{aligned}$$

From (3.8), we have

$$\begin{aligned}
\sum_{i=1}^{\ell} (\nu_i - \gamma' z_i)^2 + (\gamma - \gamma_0)' \Delta_0^{-1} (\gamma - \gamma_0) &= \sum_{i=1}^{\ell} (\nu_i - \hat{\gamma}' z_i)^2 + (\hat{\gamma} - \gamma)' \hat{\Delta}^{-1} (\hat{\gamma} - \gamma) \\
& \quad + (\gamma - \gamma_0)' \Delta_0^{-1} (\gamma - \gamma_0), \tag{E.1}
\end{aligned}$$

where $\hat{\gamma} = (z'z)^{-1}(z'\nu)$ and $\hat{\Delta} = (z'z)^{-1}$. We further simplify the right side of (E.1). First,

$$\begin{aligned}
\sum_{i=1}^{\ell} (\nu_i - \hat{\gamma}' z_i)^2 &= (\nu - z\hat{\gamma})' (\nu - z\hat{\gamma}) \\
&= \nu' \nu - 2\hat{\gamma}' z' \nu + \hat{\gamma}' z' z \hat{\gamma} \\
&= \nu' \nu - 2\nu' z (z'z)^{-1} z' \nu + \nu' z (z'z)^{-1} z' z (z'z)^{-1} z' \nu \quad [\text{since } \hat{\gamma} = (z'z)^{-1}(z'\nu)] \\
&= \nu' \nu - 2\nu' z (z'z)^{-1} z' \nu + \nu' z (z'z)^{-1} z' \nu \\
&= \nu' \nu - \nu' z (z'z)^{-1} z' \nu \\
&= \nu' [I - z(z'z)^{-1} z'] \nu. \tag{E.2}
\end{aligned}$$

Second, because $(\underline{c} - \underline{a})'A(\underline{c} - \underline{a}) + (\underline{c} - \underline{b})'B(\underline{c} - \underline{b}) = (\underline{c} - \underline{c})'(A + B)(\underline{c} - \underline{c})$
 $+ (\underline{a} - \underline{b})'A(A + B)^{-1}B(\underline{a} - \underline{b}),$

with $\underline{c} = (A + B)^{-1}(Aa + Bb)$, where A and B are symmetric matrices and A^{-1} , B^{-1} and $(A + B)^{-1}$ exist, we can write

$$(\gamma - \hat{\gamma})'\hat{\Delta}^{-1}(\gamma - \hat{\gamma}) + (\gamma - \gamma_0)'\Delta_0^{-1}(\gamma - \gamma_0) = [\gamma - (\hat{\Delta}^{-1} + \Delta_0^{-1})^{-1}(\hat{\Delta}^{-1}\hat{\gamma} + \Delta_0^{-1}\gamma_0)]'$$

$$(\hat{\Delta}^{-1} + \Delta_0^{-1})^{-1}[\gamma - (\hat{\Delta}^{-1} + \Delta_0^{-1})^{-1}(\hat{\Delta}^{-1}\hat{\gamma} + \Delta_0^{-1}\gamma_0)]$$

$$+ (\hat{\gamma} - \gamma_0)'\hat{\Delta}^{-1}(\hat{\Delta}^{-1} + \Delta_0^{-1})^{-1}\Delta_0^{-1}(\hat{\gamma} - \gamma_0).$$

Using $\Delta_0 = \kappa\hat{\Delta}$ so that $\Delta_0^{-1} = \frac{1}{\kappa}\hat{\Delta}^{-1}$, we get

$$(\hat{\Delta}^{-1} + \Delta_0^{-1})^{-1}(\hat{\Delta}^{-1}\hat{\gamma} + \Delta_0^{-1}\gamma_0) = \frac{1}{\kappa + 1}(\kappa\hat{\gamma} + \gamma_0)$$

and

$$\hat{\Delta}^{-1}(\hat{\Delta}^{-1} + \Delta_0^{-1})^{-1}\Delta_0^{-1} = \frac{1}{\kappa + 1}\hat{\Delta}^{-1}.$$

Therefore,

$$(\gamma - \hat{\gamma})'\hat{\Delta}^{-1}(\gamma - \hat{\gamma}) + (\gamma - \gamma_0)'\Delta_0^{-1}(\gamma - \gamma_0) = \left(\gamma - \frac{\kappa\hat{\gamma} + \gamma_0}{\kappa + 1}\right)' \left(\frac{\kappa + 1}{\kappa}\right) \hat{\Delta}^{-1} \left(\gamma - \frac{\kappa\hat{\gamma} + \gamma_0}{\kappa + 1}\right)$$

$$+ (\hat{\gamma} - \gamma_0)'\frac{\hat{\Delta}^{-1}}{\kappa + 1}(\hat{\gamma} - \gamma_0). \quad (\text{E.3})$$

Here,

$$(\hat{\gamma} - \gamma_0)'\frac{\hat{\Delta}^{-1}}{\kappa + 1}(\hat{\gamma} - \gamma_0) = \hat{\gamma}'\frac{\hat{\Delta}^{-1}}{\kappa + 1}\hat{\gamma} - 2\gamma_0'\frac{\hat{\Delta}^{-1}}{\kappa + 1}\hat{\gamma} + \gamma_0'\frac{\hat{\Delta}^{-1}}{\kappa + 1}\gamma_0$$

$$= \nu'z(z'z)^{-1}\frac{\hat{\Delta}^{-1}}{\kappa + 1}(z'z)^{-1}z'\nu - 2\gamma_0'\frac{\hat{\Delta}^{-1}}{\kappa + 1}(z'z)^{-1}z'\nu + \gamma_0'\frac{\hat{\Delta}^{-1}}{\kappa + 1}\gamma_0$$

$$= \nu'z(z'z)^{-1}\frac{(z'z)}{\kappa + 1}(z'z)^{-1}z'\nu - 2\gamma_0'\frac{(z'z)}{\kappa + 1}(z'z)^{-1}z'\nu + \gamma_0'\frac{(z'z)}{\kappa + 1}\gamma_0, \text{ [since, } \hat{\Delta} = (z'z)^{-1}]$$

$$= \nu'z\frac{(z'z)^{-1}}{\kappa + 1}z'\nu - 2\gamma_0'\frac{(z'\nu)}{\kappa + 1} + \gamma_0'\frac{(z'z)}{\kappa + 1}\gamma_0.$$

Therefore,

$$\begin{aligned}
(\gamma - \hat{\gamma})' \hat{\Delta}^{-1} (\gamma - \hat{\gamma}) + (\gamma - \gamma_0)' \Delta_0^{-1} (\gamma - \gamma_0) &= \left(\gamma - \frac{\kappa \hat{\gamma} + \gamma_0}{\kappa + 1} \right)' \left(\frac{\kappa + 1}{\kappa} \right) \hat{\Delta}^{-1} \left(\gamma - \frac{\kappa \hat{\gamma} + \gamma_0}{\kappa + 1} \right) \\
&\quad + \nu' z \frac{(z'z)^{-1}}{\kappa + 1} z' \nu - 2\gamma_0' \frac{(z'\nu)}{\kappa + 1} + \gamma_0' \frac{(z'z)}{\kappa + 1} \gamma_0.
\end{aligned} \tag{E.4}$$

Now, using (E.2) and (E.3) in (E.1), we get

$$\begin{aligned}
&\sum_{i=1}^{\ell} (\nu_i - \gamma' z_i)^2 + (\gamma - \gamma_0)' \Delta_0^{-1} (\gamma - \gamma_0) \\
&= \nu' [I - z(z'z)^{-1} z'] \nu + \frac{1}{\kappa + 1} \nu' z (z'z)^{-1} z' \nu - \frac{2}{\kappa + 1} \gamma_0' z' \nu + \gamma_0' \frac{(z'z)}{\kappa + 1} \gamma_0 \\
&\quad + \left(\gamma - \frac{\kappa \hat{\gamma} + \gamma_0}{\kappa + 1} \right)' \left(\frac{\kappa + 1}{\kappa} \right) \hat{\Delta}^{-1} \left(\gamma - \frac{\kappa \hat{\gamma} + \gamma_0}{\kappa + 1} \right) \\
&= \gamma_0' \frac{(z'z)}{\kappa + 1} \gamma_0 + \nu' [I - \frac{\kappa}{\kappa + 1} z(z'z)^{-1} z'] \nu - \frac{2}{\kappa + 1} \gamma_0' z' \nu \\
&\quad + \left(\gamma - \frac{\kappa \hat{\gamma} + \gamma_0}{\kappa + 1} \right)' \left(\frac{\kappa + 1}{\kappa} \right) \hat{\Delta}^{-1} \left(\gamma - \frac{\kappa \hat{\gamma} + \gamma_0}{\kappa + 1} \right).
\end{aligned} \tag{E.5}$$

Appendix E

Proof that $\int_{-\infty}^{\infty} h(\nu_i) d\nu_i$ is finite.

From (3.17), we have

$$\begin{aligned}
F &= \int_{\phi} \int_{\underline{y}^{(1)}} \int_{\sigma^2} g(\phi) B(\phi) (1/\sigma^2)^{(\ell+a)/2-1} \\
&\quad \times e^{-\frac{1}{2\sigma^2} [b + \underline{y}^{(1)'} \underline{y}^{(1)}]} \prod_{i=\ell-q+1}^{\ell} \left\{ \int_{-\infty}^{\infty} h(\nu_i) d\nu_i \right\} d\sigma^2 d\underline{y}^{(1)} d\phi.
\end{aligned}$$

where $h(\nu_i) = \frac{e^{\nu_i a_i}}{\prod_{j=1}^{n_i} [1 + e^{\nu_i c_{ij}]}$ as defined in (3.14) with $0 \leq a_i \leq n_i$ and $c_{ij} \geq 0$. It is difficult to find the exact definite integral of $h(\nu_i)$ because of the complicated integrand. However, we can show that $\int_{-\infty}^{\infty} h(\nu_i) d\nu_i$ is finite by bounding the integrand with something simpler.

Let $n_i = n$ and then dropping subscript i and assuming that all c_j are same (equal to c),

we have from (G.1)

$$h(\nu_i) = \frac{e^{\nu a}}{\prod_{j=1}^n (1 + e^{\nu c})^n}.$$

Now,

$$\int_{-\infty}^{\infty} h(\nu) d\nu = \int_{-\infty}^0 \frac{e^{\nu a}}{\prod_{j=1}^n (1 + e^{\nu c})^n} d\nu + \int_0^{\infty} \frac{e^{\nu a}}{\prod_{j=1}^n (1 + e^{\nu c})^n} d\nu. \quad (\text{G.1})$$

Note here in the first part of (G.1), $1 + e^{\nu c} \geq 1$ for ν in $[-\infty, 0]$ and $c \geq 0$. Therefore

$$\frac{e^{a\nu}}{(1 + ce^{\nu})^n} \leq e^{a\nu}.$$

Similarly, on the second part of (G.1)

$$\frac{e^{a\nu}}{(1 + ce^{\nu})^n} \leq \frac{e^{a\nu}}{c^n e^{n\nu}} = (1/c^n) e^{-(n-a)\nu}.$$

Thus, from (G.1)

$$\int_{-\infty}^{\infty} h(\nu) d\nu \leq \int_{-\infty}^0 e^{\nu a} d\nu + \int_0^{\infty} (1/c^n) e^{-(n-a)\nu} d\nu.$$

Because $n \geq a$, we get

$$\int_{-\infty}^{\infty} h(\nu) d\nu = \frac{1}{a} + \frac{1}{c^n(n-a)} < \infty.$$

Appendix F

Cluster Tables for Examples E1-E6 using TIMSS 2007 Data

Clusters	n	<u>E1</u>			
		(1,1)	(1,2)	(2,1)	(2,2)
1	9	1	1	0	7
2	48	20	9	9	10
3	21	0	2	2	17
4	39	2	6	5	26
5	21	6	3	1	11
6	44	20	6	7	11
7	41	0	0	4	37
8	38	23	3	3	9
9	44	10	7	3	24
10	35	15	7	5	8
11	35	6	7	3	19
12	25	12	2	5	6
13	32	5	8	6	13
14	33	8	6	4	15
15	39	13	4	8	14
16	12	9	1	1	1
17	21	7	5	5	4
18	28	7	6	4	11
19	31	3	4	2	22
20	21	16	2	1	2
21	32	1	1	2	28
22	29	12	8	2	7
23	19	9	3	3	4
24	34	16	3	8	7
25	27	4	8	4	11
26	23	4	4	4	11

Clusters	n	<u>E2</u>			
		(1,1)	(1,2)	(2,1)	(2,2)
1	19	3	2	1	13
2	14	1	1	0	12
3	34	12	9	7	6
4	35	3	0	2	30
5	23	0	2	4	17
6	43	8	7	2	26
7	41	1	1	2	37
8	35	26	3	2	4
9	46	9	8	12	17
10	37	7	2	7	21
11	34	17	7	5	5
12	36	17	4	10	5
13	29	17	2	4	6
14	40	18	4	9	9
15	25	13	4	3	5
16	15	7	4	0	4
17	29	2	4	1	22
18	26	6	4	3	13
19	34	2	2	1	29
20	27	5	2	7	13
21	10	7	1	0	2
22	29	2	4	4	19
23	39	14	4	3	18
24	32	15	4	6	7
25	41	9	8	13	11
26	43	7	4	7	25
27	21	11	4	2	4

Clusters	n	<u>E3</u>			
		(1,1)	(1,2)	(2,1)	(2,2)
1	18	12	0	2	4
2	29	16	4	3	6
3	36	14	5	6	11
4	32	15	2	4	11
5	44	5	5	5	29
6	32	19	5	4	4
7	34	24	2	2	6
8	24	19	3	1	1
9	35	2	4	2	27
10	42	17	4	8	13
11	32	12	7	6	7
12	36	3	4	3	26
13	35	13	6	5	11
14	29	7	2	3	17
15	24	3	4	4	13
16	21	6	4	3	8
17	14	4	2	5	3
18	25	4	3	2	16
19	34	2	2	3	27
20	37	12	6	1	18
21	24	0	2	1	21
22	21	17	0	3	1
23	25	4	7	2	12
24	4	1	0	0	3
25	31	1	2	2	26
26	26	7	3	5	11

Clusters	n	<u>E6</u>			
		(1,1)	(1,2)	(2,1)	(2,2)
1	11	3	5	1	2
2	17	5	5	2	5
3	10	9	0	1	0
4	20	6	5	3	6
5	31	4	8	3	16
6	31	0	2	1	28
7	24	6	5	0	13
8	12	2	1	1	8
9	19	5	1	4	9
10	39	5	7	1	26
11	32	12	4	3	13
12	23	4	4	1	14
13	34	9	8	3	14
14	29	4	5	4	16
15	19	15	2	1	1
16	28	10	3	5	10
17	26	5	8	3	10
18	26	4	5	3	14
19	21	4	8	1	8
20	12	2	5	0	5
21	28	2	5	0	21
22	6	0	2	0	4
23	32	12	3	7	10
24	29	5	6	3	15
25	16	8	5	2	1

Clusters	n	E5			
		(1, 1)	(1, 2)	(2, 1)	(2, 2)
1	25	8	5	2	10
2	30	5	1	7	17
3	27	3	5	4	15
4	25	4	4	2	15
5	23	5	7	4	7
6	34	7	5	7	15
7	40	8	4	4	24
8	32	11	6	4	11
9	26	2	6	4	14
10	35	6	4	5	20
11	42	6	9	6	21
12	32	5	7	2	18
13	22	5	5	1	11
14	26	14	2	7	3
15	32	10	9	2	11
16	30	1	2	1	26
17	35	11	6	3	15
18	31	10	2	3	16
19	21	1	3	1	16
20	26	5	6	2	13
21	34	5	2	3	24
22	41	0	6	3	32
23	34	10	9	4	11
24	26	17	7	2	0
25	28	20	3	4	1
26	19	5	3	4	7
27	21	4	4	4	9
28	32	6	9	4	13
29	21	0	4	1	16
30	34	5	3	3	23
31	33	3	8	5	17
32	27	10	4	3	10
33	29	5	1	8	15
34	29	13	9	1	6
35	31	7	1	6	17
36	27	3	3	4	17
37	21	3	9	1	8
38	34	9	16	1	8
39	31	3	11	0	17
40	22	8	5	2	7
41	26	12	3	3	8
42	28	2	5	4	17
43	26	3	6	1	16
44	30	5	6	7	12
45	26	13	4	4	5
46	29	2	3	6	18
47	26	3	5	2	16
48	31	2	6	5	18
49	41	20	9	8	4
50	9	0	0	1	8
51	8	3	1	1	3
52	28	21	1	5	1
53	27	10	3	8	6
54	30	5	6	2	17
55	26	12	4	5	5
56	24	10	4	2	8
57	36	9	10	4	13
58	23	9	9	1	4
59	33	8	5	2	18

Clusters	n	E4			
		(1, 1)	(1, 2)	(2, 1)	(2, 2)
1	26	2	3	3	18
2	37	3	8	4	22
3	27	8	7	5	7
4	36	6	6	7	17
5	32	1	6	2	23
6	38	1	8	3	26
7	38	1	2	3	32
8	31	18	5	5	3
9	36	18	4	5	9
10	30	8	4	3	15
11	23	8	6	1	8
12	32	7	5	10	10
13	42	7	11	3	21
14	32	5	8	3	16
15	29	3	3	6	17
16	19	4	5	1	9
17	37	3	2	6	26
18	29	2	4	4	19
19	33	1	5	2	25
20	27	2	5	2	18
21	32	5	2	2	23
22	26	13	4	3	6
23	38	0	2	3	33
24	28	3	3	3	19
25	32	8	10	2	12
26	28	13	6	0	9
27	36	5	3	4	24
28	34	3	4	3	24
29	25	3	11	2	9
30	14	2	4	0	8
31	16	3	1	4	8
32	23	0	3	2	18
33	22	3	3	4	12
34	24	5	3	4	12
35	36	8	4	2	22
36	27	8	2	6	11
37	33	19	2	6	6
38	33	9	3	3	18
39	31	5	4	2	20
40	36	6	7	4	19
41	30	4	6	6	14
42	22	5	4	3	10
43	37	10	6	6	15
44	29	4	5	3	17
45	26	4	4	2	16
46	35	0	6	1	28
47	24	3	2	4	15
48	32	3	8	4	17
49	24	6	6	2	10

References

- Arnold, B. C. and Strauss, D. (1991), "Pseudolikelihood Estimation: Some Examples," *Sankhya: The Indian Journal of Statistics, Series B*, 53(2), 233-243.
- Bedrick, E. J. (1983), "Chi-Squared Tests for Cross-classified Tables of Survey Data," *Biometrika*, 70(3), 591-595.
- Bhatta, D. and Nandram, B., "A Bayesian Adjustment of the HP law Using a Switching Nonlinear Regression Model," *Journal of Data Science*, 11(2013), 85-108.
- Brier, S. S. (1980), "Analysis of Contingency Tables Under Cluster Sampling," *Biometrika*, 67(3), 591-596.
- Fellegi, I. P. (1980), "Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples," *Journal of the American Statistical Association*, 75, 261-268.
- Geenens, G. and Simar, L. (2010), "Nonparametric tests for conditional independence in two-way contingency tables," *Journal of Multivariate Analysis*, 101, 765-788.
- Geenens, G. and Simar, L. (2011), "Single-index modelling of conditional probabilities in two-way contingency tables," *A Journal of Theoretical and Applied Statistics*, 45(5), 451478.
- Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004), "Bayesian Data Analysis (2nd Edition)," *New York: Chapman & Hall/CRC*.
- Gelman, A. (2007), "Struggles with Survey Weighting and Regression Modeling," *Statistical Science*, 22(2), 153-164.
- Gilks, W. R. and Wild, P. (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(2), 337-348.
- Gonzalez, E. J. and Smith, T.A. (1997), "Users Guide for the TIMSS International Database," *Chestnut Hill, MA: TIMSS International Study Center*.
- Hjort, N.L., Dahl, F. A. and Steinbakk, G. H. (2006), "Post-Processing Posterior Predictive p Values," *Journal of the American Statistical Association*, 101(475), 1157-1174.
- Kass, R.E. and Raftery, A.E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90(430), 773-795.
- Malec, D., Davis, W. W. and Cao, X. (1999), "Model-based Small Area Estimates of Overweight Prevalence Using Sample Selection Adjustment," *Statistics in Medicine*, 18, 3189-3200.
- Nandram, B. and Choi, J.W. (2007), "Alternative Tests of Independence in Two-Way Categorical Tables," *Journal of Data Science*, 5, 217-237.
- Nandram, B. (2007), "Bayesian Predictive Inference Under Informative Sampling via Surrogate Samples," *Bayesian Statistics and Its Applications*, Eds. S.K. Upadhyay, Umesh Singh and Dipak K. Dey, Anamaya, New Delhi, Chapter 25, 356-374.

- Nandram, B. and Sedransk, J. (1993), "Bayesian Predictive Inference for a Finite Population Proportion: Two-Stage Cluster Sampling," *Journal of the Royal Statistical Society, Series B*, 55(2), 399-408.
- Nandram, B. (1998), "A Bayesian Analysis of the Three-stage Hierarchical Multinomial Model," *Journal of Statistical Computation and Simulation*, 61, 97-126.
- Nandram, B. and Choi, J. W. (2010), "A Bayesian Analysis of Body Mass Index Data from Small Domains Under Nonignorable Nonresponse and Selection," *Journal of the American Statistical Association*, 105, 120-135.
- Nandram, B., Bhatta, D. and Bhadra, D. (2012), "A Likelihood Ratio Test of Quasi-Independence for Sparse Two-Way Contingency Tables," *Journal of Statistical Computation and Simulation* (under review).
- Nandram, B., Bhatta, D., Bhadra, D. and Shen, G. (2012), "Bayesian Predictive Inference of a Finite Population Proportion Under Selection Bias," *Statistical Methodology*, 11 (2013) 121.
- Nandram, B., Bhatta, D., Bhadra, D., and Sedransk, J., "A Bayesian Test of Independence in a Two-Way Contingency Table Using Surrogate Sampling," *Journal of Statistical Planning and Inference* (under review).
- Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H. and Rasbash, J. (1998), "Weighting for Unequal Selection Probabilities in Multilevel Models," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60(1), 23-40.
- Rao, J. N. K. and Scott, A. J. (1981), "The Analysis of Categorical Data From Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables," *Journal of the American Statistical Association*, 76(374), 221-230.
- Rao, J. N. K. and Scott, A. J. (1984), "On Chi-Squared Tests for Multiway Contingency Tables with Cell Proportions Estimated From Survey Data," *The Annals of Statistics*, 12(1), 46-60.
- Rao, J. N. K. and Thomas, D. R. (1989), "Chi-squared Tests for Contingency Tables," *In the Analysis of Complex Surveys*, Eds. D. Holt, C.J. Skinner and T.M.F. Smith, New York: Wiley.
- Ritter, C. and Tanner, M.A. (1992), "Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler," *Journal of the American Statistical Association*, 87(419), 861-868.
- Robert, C. P. and Casella, G. (1999), *Monte Carlo Statistical Methods*, New York: Springer.
- Satterthwaite, F.E. (1946), "An Approximate Distribution of Estimates of Variance Components," *Biometrics Bulletin*, 2(6), 110-114.
- Scott, A. (2007), "Rao-Scott Corrections and Their Impact," *Section on Survey Research Methods-JSM*, 3514-3518.

Seber, G. A. F. (1984), "Multivariate Observations," *Wiley Series in Probability and Mathematical Statistics*.

Silverman, B. W. (1986), "Density Estimation for Statistics and Data Analysis," New York: Chapman & Hall.

Thomas, D. R. and Rao, J. N. K. (1987), "Small Sample Comparisons of Level and Power for Simple Goodness of Fit Statistics Under Cluster Sampling," *Journal of the American Statistical Association*, 82, 630-636.

Thomas, D. R., Singh, A. C. and Roberts, G. R. (1996), "Tests of Independence on Two-Way Tables under Cluster Sampling: An Evaluation," *International Statistical Review / Revue Internationale de Statistique*, 64(3), 295-311.

"TIMSS 2007 U.S. Technical Report and User Guide,"
(nces.ed.gov/pubs2009/2009012_2.pdf.)

Wang, J. (2005), "Relationship Between Mathematics and Science Achievement at the 8th Grade," *Online Submission, International Journal of Science and Math Education*, 5, 1-17.

Wei, R., Nandram, B. and Bhatta, D., "A Bayesian Analysis of US Mortality Curves for Race-Sex Domains by State," *Statistics in Medicine* (under review).

Zelterman, D. (1987), "Goodness-of-Fit Tests for Large Sparse Multinomial Distributions," *Journal of the American Statistical Association*, 82(398), 624-629.