International Journal of Statistical Sciences Vol. 9(Special Issue), 2009, pp 273-284 © 2009 Dept. of Statistics, Univ. of Rajshahi, Bangladesh

## A Semi-Bayesian Approach for Estimation of Joint Distribution from Marginal Distributions

#### **Yogendra P. Chaubey**

Department of Mathematics and Statistics, Concordia University, 1455 de Maisonneuve Blvd. West, Montreal, QC, H3G 1M8, Canada E-mail: chaubey@alcor.concordia.ca

### Fassil Nebebe

Department of Decision Sciences and MIS, Concordia University, 1455 de Maisonneuve Blvd. West, Montreal, QC, H3G 1M8, Canada E-mail: fnebebe@jmsb.concordia.ca

#### Krzysztof Dzieciolowski

*Fido Solutions* 800 de La Gauchetiere West #4000, Montreal, QC, H5A 1K3, Canada, E-mail: kdzieciolowski@jmsb.concordia.ca

#### Debaraj Sen

Department of Mathematics and Statistics, Concordia University, 1455 de Maisonneuve Blvd. West, Montreal, QC, H3G 1M8, Canada E-mail: sen@mathstat.concordia.ca

[Received January 21, 2009; Revised February 17, 2009; Accepted February 25, 2009]

### Abstract

In many consumer surveys, demographic and other data are used as covariates for predictions of consumer preferences and other decision variables. However, to protect the confidentiality of the consumers, the data are only available in marginal frequency distribution format. This creates a problem in predicting joint frequencies required for further decision-making. In this paper we consider the dependence structure in formulating the joint distribution given in the form of a Dirichlet prior. The computations in the case of  $2 \times 2$  table are extended to multiway contingency tables and shown to provide similar results as obtained using the Monte Carlo methods.

Keywords and Phrases: Bayesian Prediction, Contingency Table, Dirichlet Prior.

AMS Classification: Primary 62F15; Secondary 62E99.

ISSN 1683–5603

## **1** Introduction

In this paper we consider estimating the joint distribution of several demographic variable when only information about their marginal distributions is available. Putler *et al.* (1996) considered a Bayesian approach using Dirichlet prior for the cell proportions. They also compared iterative proportional fitting (IPF) method (see Bishop, Feinberg, and Holland 1975). They presented the case of  $2 \times 2$  table in detail. However, the higher contingency tables have to be analyzed using Monte Carlo methods, because they require multidimensional numerical integration. However, the Bayesian method is preferred as it readily presents an estimate of the standard error derived from the posterior variance.

The proposal in this paper is to reduce the contingency table into several  $2 \times 2$  contingency tables, estimate the target cell frequency along with its standard error and combine the results obtained from various tables in an appropriate manner. We highlight here that computations in the case of  $2 \times 2$  table can be easily performed. For example, suppose, we have a  $2 \times 2 \times 2$  contingency table, with unknown cell proportions written as  $x_{ijk}$ ; i = 1, 2, j = 1, 2, k = 1, 2. For estimating  $x_{111}$ , we can consider the following  $3, 2 \times 2$  contingency tables:

	k = 1					
		j = 1	j = 2			
i = 1	1	$x_{111}$	$x_{121}$			
i = 2	2	$x_{211}$	$x_{221}$			

Table 1: Three 2  $\times$  2 Contingency Tables for Estimating  $x_{111}$  from a 2  $\times$  2  $\times$  2 Table

	i = 1					
		k = 1	k = 2			
j = 1	1	$x_{111}$	$x_{112}$			
j = 2	2	$x_{121}$	$x_{122}$			

		j = 1				
		k = 1	k = 2			
i = 1	1	$x_{111}$	$x_{112}$			
i = 2	2	$x_{211}$	$x_{212}$			

Let,  $\hat{x}_{111}^{(l)}$  and  $se^{2(l)}(\hat{x}_{111})$  denote the posterior mean and posterior variance based on the  $l^{\text{th}}$  table comprising of  $r_l$  observation, the final estimates of  $x_{111}$  and its proposed estimator of error is given by;

$$\hat{x}_{111} = \frac{\sum_{l=1}^{3} r_l \hat{x}_{111}^{(l)}}{\sum_{l=1}^{3} r_l}$$
(1.1)

$$se^{2}(x_{111}) = \frac{\sum_{l=1}^{3} r_{l} se^{2(l)}(\hat{x}_{111})}{\sum_{l=1}^{3} r_{l}}$$
(1.2)

In section 2, we present the Bayesian analysis of a  $2 \times 2$  contingency table as given in Putler *et al.* (1996). Section 3 outlines the computational details along with an example. Section 4 presents computational summary of the examples considered in Putler *et al.* (1996) and some additional examples from a marketing consumer survey data in Montréal.

# **2** The Bayesian Approach in $2 \times 2$ Case

Consider as if we have the observed data in the form of a  $2 \times 2$  contingency table:

Table 2: A Typical  $2 \times 2$  Contingency Table

	Column Factor Levels				
		1	2	Marginal	
Row	1	$nx_{11}$	$nx_{12}$	$nx_{1.}$	
Factor					
Levels	2	$nx_{21}$	$nx_{22}$	$nx_{2.}$	
	Marginal	$nx_{.1}$	$nx_{.2}$	$nx_{} = n$	

where

$$x_{ij} \in [0,1], \sum_{ij} x_{ij} = 1,$$

and n is the sample size. The set-up we are concerned with is that  $x_{ij}$  are unknown, however the row-totals and column-totals are known. Of course, under some given assumptions such as independence, the known marginal probabilities can be used to estimate the joint probabilities

$$p_{ij} = \Pr[X_1 = i, X_2 = j],$$

where  $X_1$  refers to the row attribute and  $X_2$  refers to the column attribute. Putler *et al.* (1996) consider a joint Dirichlet prior on  $p = (p_{11}, p_{12}, p_{21}, p_{22})$  given by

$$\pi(p) \propto \prod_{ij} p_{ij}^{m\alpha_{ij}-1}$$

where  $\alpha_{ij}$  refer to the prior information about  $p_{ij}$ , often available from some bench-mark data or a larger survey, so that

$$\alpha_{ij} \in [0,1], \ \sum_{ij} \alpha_{ij} = 1,$$

and m refers to the weight assigned to the prior information. The prior implicitly assumes that in m observations of prior data, the total in the (i, j)-cell is  $m\alpha_{ij}$ .

With this set-up, because of the constraints on the cell-proportions, only one of the cell-totals is independent. We choose arbitrarily the (2, 2) - cell, and define the random variable  $z = nx_{22}$ . The posterior inference about z may be based on the density g(z|data), given by

$$g(z) = k \{ \Gamma(n_{1.} - n_{.2} + z + m\alpha_{11}) \Gamma(n_{.2} - z + m\alpha_{12})$$

$$\Gamma(n_{2.} - z + m\alpha_{21}) \Gamma(z + m\alpha_{22}) /$$

$$\Gamma(n_{1.} - n_{.2} + z + 1) \Gamma(n_{.2} - z + 1) \Gamma(n_{2.} - z + 1) \Gamma(z + 1) \},$$
(2.1)

where  $n_{i.} = nx_{i.}$ ,  $n_{.j} = nx_{.j}$ ; i, j = 1, 2. Putler *et al.* (1996) suggest that the posterior mean can be computed in a two step procedure. The first step consists of computing the constant k of proportionality, through a trapezoidal rule over a large number of intervals. The second step consists of evaluating the integral

$$\hat{z} = E(Z|data) = \int_{z_{min}}^{z_{max}} zg(z)dz,$$

again using the trapezoidal rule, where

$$z_{min} = \max(0, n_{.2} - n_{1.}), \ z_{max} = \min(n_{.2}, n_{2.}).$$

An estimate of the standard error of  $\hat{z}$  is obtained from the posterior variance given by

$$se^{2}(Z) = E[(Z - \hat{z})^{2}|data] = \int_{z_{min}}^{z_{max}} (z - \hat{z})^{2}g(z)dz$$

which is computed using the same approach as that for computing  $\hat{z}$ .

## **3** Computational Aspects

For the general case, Monte Carlo (MC) methods are proposed for estimation of proportions  $p_{i_1i_2...}$ , such as Importance Sampling and Gibbs Sampling. It is clear that MC methods are very computationally intensive in this context and therefore, we investigate the possibility of direct computation as suggested in the previous section. In practice, though the computation of gamma functions arising in Eq. (1.1) presents problems. In most of the situations, we may alleviate this problem by using the following technique. Let us denote  $z_0$ , the approximate value of z where g(z) takes its maxima (it is shown in Putler *et al.* (1996) that g(z) is unimodal). We can thus write

$$E(Z|data) = \frac{\int_{z_{min}}^{z_{max}} z \exp(h(z) - h(z_0)) dz}{\int_{z_{min}}^{z_{max}} \exp(h(z) - h(z_0)) dz},$$
(3.1)

where, h(z) is defined, so that

 $g(z) = k \exp(h(z)).$ 

The value of  $z_0$  can be found by plotting h(z) against z, which may be computable, where as exp(h(z)) may not be. We first demonstrate this on the data used in Putler *et al.* (1996), and then we use it on a survey data conducted by a market research firm. For the computation of the integral, we use the following R-codes for area function.

```
area<-function(f,a,b,...,fa=f(a,...),fb=f(b,...),limit=50,eps=1.0e-06)
    {
#Program to integrate a function f using recursive Simpson's rule
#eps is the absolute target error #limit is max number of iterations
   h<- b-a
   d<- (b+a)/2
    fd<- f(d,...)
    a1 < -((fa+fb)*h)/2
    a2<-((fa+4*fd+fb)*h)/6
    if (abs(a1-a2) < eps)
       return(a2)
        if (limit ==0){
            warning(paste("recursion limit reached near x= ",d))
            return(a2)
                      }
     Recall(f,a,d,...,fa=fa,fb=fd,limit=limit-1,eps=eps)+
        Recall(f,d,b,...,fa=fd,fb=fb,limit=limit-1,eps=eps)
    }
```

This function works pretty well, except for the cases where the function f may take extremely small values for a wide range of arguments. For example, if f(x) < eps for x in an interval (c, b), c < (a + b)/2, and f(a) = f(b) = 0, the algorithm will produce a small but wrong value of the integral. To avoid such situations, we evaluate the integral over two intervals,  $(a, z_0)$  and  $(z_0, b)$ , where  $z_0$  is the approximate value of the argument where the function peaks. The approximate peak is obtained by taking the maximum of h(z) evaluated over a grid of z-values.

**Example:** Consider the data of Putler *et al.* (1996) on Stain-Resistent Carpeting Direct Mail Campaign (see their Table 5). We consider a collapsed form of the contingency table into two factors, Type of Housing and Marital Status of the Household Head. The actual proportions are given in the following table (Note that the actual proportions in Table 5 of Putler *et al.* (*loc. cit.*) do not add to 1 so we have arbitrarily chosen to adjust the last proportion from .2520 to .2519).

	Household Status							
		Total						
Type of	Rental	.2670	.0874	.3544(4552)				
Housing	Owned	.1661	.4795	.6456(8291)				
	Total	.4331(5562)	.5669(7281)	n = 12843				

Table 3: Proportions in a  $2 \times 2$  Contingency Table

The numbers in the parentheses are counts. The prior-probabilities are given in the following table and the value of the weight m is given by m = 12843:

	Household Status						
		Total					
Type of	Rental	.3072	.0955	.3544			
Housing	Owned	.1793	.4180	.6456			
	Total	.4331	.5669				

Table 4: Prior Proportions for Contingency Table 3

The following R-Codes are used to compute the function h and approximate  $z_0$ .

278

```
h.prior<-function(x, rowt, colt, pprob, pcount)
  #log Density-kernel for Estimating Unknown Proportions
  #this finds the prior for a cell count,
        see Eq (6) Putler et al.
  # lam is the variable in the (2,2) cell
  # pprob is the matrix of prior probabilities
  # pcount is the weight in the Dirichlet prior,
        see Eq (2) of Putler et al.
  # rowt is a vector of row totals
  # colt is a vector of column totals
  # definition of the probability kernel
 x1. <- rowt[1]
 x2. <- rowt[2]
 x.1 <- colt[1]
 x.2 <- colt[2]
 n < -x1. + x2.
 all <-x1. -x.2 + x + pcount * pprob[1,1]
 al2 <- x.2 - x + pcount * pprob[1, 2]
 b11 < x2. - x + pcount * pprob[2, 1]
 b12 <- x + pcount * pprob[2, 2]
 cl1 <- x1. - x.2 + x + 1.
  c12 < -x.2 - x + 1
 d11 <- x2. - x + 1.
 d12 <- x + 1.
 tnum <- c(all, al2, bl1, bl2)
  tden <- c(c11, c12, d11, d12)
  sum(lgamma(tnum))-sum(lgamma(tden)) }
```

The value of  $z_{min}$  and  $z_{max}$  are  $z_{min} = 7281 - 4552 = 2729$ ,  $z_{max} = 7281$  and the values of the variables rowt, colt, pprob, pcount are set as follows:

```
rowt<-c(4552, 8291)
colt<-c(5562, 7281)
pprob<-matrix(c(0.2670, 0.0874, 0.1661, 0.4795),nr=2,byrow=T)
pcount<- 12843</pre>
```

The graph of the function h(z) is then obtained as follows and appears in Figure 1.

```
zseq<-seq(2729,7281,length=20) hzseq<-sapply(zseq,h.prior,
rowt=rowt, colt=colt, pprob=pprob, pcount=pcount)
plot(zseq,hzseq)
```



Figure 1: Plot of the Function h(z)

This gives the value of  $z_0$  approximately 6000. So we compute  $h(z_0)$  as

```
>h.prior(6000,rowt,colt,pprob,pcount)
[1] 110799.4
```

so,  $h(z_0) = 110799.4$  and hence the function g(z) can be defined as

```
g.prior<-function(x, rowt, colt, pprob, pcount) {
        exp(h.prior(x, rowt, colt, pprob, pcount)-110799)
    }</pre>
```

and the denominator of Eq. (3.1) is obtained using the area over subintervals (2729, 6000) and (6000, 7281) as follows:

```
> area(g.prior,2729,6000,rowt, colt, pprob, pcount)
[1] 10.02941
> area(g.prior,6000,7281,rowt, colt, pprob, pcount)
[1] 7382576
```

280

This comes out to be 7382586.02941. Hence the expression for the posterior mean can be obtained through the following codes:

```
> kernel.mean<-function(x, rowt, colt, pprob, pcount)x *
    + g.prior(x, rowt, colt, pprob, pcount)/7382586.03
> area(kernel.mean,2729,6000,rowt, colt, pprob, pcount)
[1] 0.008141925
> area(kernel.mean,6000,7281,rowt, colt, pprob, pcount)
[1] 6158.561
```

This gives the Bayes estimate of  $x_{22} = 6158.5691/12843 = .4795$ , which turns out to be the same as the actual proportion. We also computed the posterior mean for n = 12843 and m = 15023, and results were almost the same.

## 4 Summary of Numerical Illustrations for Higher Order Tables

### 4.1 Data from Putler et al. (1996)

Here we present the computations for the three examples considered in Tables 5-7 from Putler *et al.* (1996). The subscripts i, j, k refer to various demographic characteristics as follows:

- Stain-Resistant Carpeting Direct Mail Campaign (Table 5)
  - 0. Housing: Rent, i = 1. Own, i = 2.
  - 0. Marital Status of Household Head: Married, j = 1. Not Married, j = 2.
  - 0. Household Member Age: < 18, k = 1.  $\geq 18, k = 2$ .
- Discount Home-Improvement Retail Site Location (Table 6)
  - 0. Housing: Rent, i = 1. Own, i = 2.
  - 0. Household Income:  $\geq 40,000, j = 1.$ < 40,000, j = 2.
  - 0. Household Member Age:  $\geq 45, k = 1.$ < 45, k = 2.
- Custom-made Golf Clubs Direct Mail Campaign Data (Table 7)
  - 0. Annual Household Income: < 50,0000, i = 1.  $\geq 50,000, i = 2$ .

- 0. Sex: Male, j = 1. Female, j = 2.
- 0. Years of Formal College Education: < 4, k = 1.  $\geq 4, k = 2$ .

The following tables (Tables 5-7) give the values of the estimates of the cell proportions along with their standard errors following the method outlined above, contrasting the values obtained by Putler *et al.* (1996) using the Monte Carlo method. It is striking that our method gives almost the same estimate as obtained by the Monte Carlo Method. The estimates of the standard errors seem a bit inflated in some cases. At this point, we like to note that since, there is only one degree of freedom in a  $2 \times 2 \times 2$  table, all the standard errors should be the same. In such cases, we can average out the standard errors in an appropriate way (such as waiting by the total sample size in each partition of sub-tables). In the above examples, we can combine the posterior variances obtained from the two sub-tables and then average them. However, in more complicated set up this type of combination may not be possible. We also investigated using the estimates under independence as priors. The resulting estimates are not far from the original prior, which seem to be much closer to the actual proportions. We like to think the independence estimates as non-informative prior which get adjusted by the data through the posterior mean.

Table 5: Summary for the Results for Table 5

Cell ID	111	112	121	122	211	212	221	222
Actual	.2252	.0418	.0413	.0461	.1378	.0283	.2276	.2520
Putler	.2219	.0366	.0333	.0622	.1453	.0292	.2315	.2400
et al.	(.0053)	(.0037)	(.0037)	(.0044)	(.0052)	(.0035)	(.0054)	(.0058)
Our	.2293	.0368	.0342	.0546	.1357	.0308	.2324	.2465
Proposal	(.0064)	(.0070)	(.0060)	(.0065)	(.0057)	(.0061)	(.0054)	(.0058)
Independent	.1955	.0315	.0691	.0544	.1476	.0541	.2179	.2292
Prior	(.0070)	(.0066)	(.0066)	(.0063)	(.0066)	(.0063)	(.0064)	(.0060)

Table 6: Summary for the Results for Table 6

Cell ID	111	112	121	122	211	212	221	222
Actual	.0400	.0471	.1055	.1243	.0862	.1015	.2274	.2680
Putler	.0109	.0240	.0870	.1951	.1116	.1285	.2495	.1934
et al.	(.0024)	(.0033)	(.0052)	(.0056)	(.0058)	(.0060)	(.0060)	(.0061)
Our	.0106	.0151	.0919	.1912	.1096	.1372	.2471	.1911
Proposal	(.0058)	(.0062)	(.0060)	(.0063)	(.0064)	(.0067)	(.0064)	(.0067)
Independent	.0180	.0328	.1089	.1575	.1167	.1163	.2275	.2242
Prior	(.0063)	(.0068)	(.0065)	(.0069)	(.0068)	(.0072)	(.0069)	(.0071)

Cell ID	111	112	121	122	211	212	221	222
Actual	.3379	.3742	.0464	.0513	.0793	.0879	.0109	.0121
Putler	.3426	.3921	.0356	.0397	.0709	.0734	.0259	.0198
et al.	(.0034)	(.0036)	(.0025)	(.0024)	(.0030)	(.0030)	(.0021)	(.0018)
Our	.3418	.3919	.0367	.0393	.0703	.0751	.0257	.0195
Proposal	(.0033)	(.0032)	(.0039)	(.0038)	(.0043)	(.0041)	(.0064)	(.0059)
Independent	.3401	.3885	.0411	.0427	.0733	.0794	.0167	.0145
Prior	(.0030)	(.0029)	(.0037)	(.0035)	(.0040)	(.0038)	(.0062)	(.0057)

Table 7: Summary for the Results for Table 7

## 4.2 Survey Data

Here we present the computations for survey data. The subscripts i, j, k refer to various demographic characteristics as follows:

- Gender: Male, i = 1. Female, i = 2.
- Age between 1 to 44 : *j* = 1. between 45 and above , *j* = 2.
- Education: less than Graduate, k = 1. Graduate and above, k = 2.

We investigated using the estimates under independence as priors. The resulting estimates are not far from the actual proportions. We like to think the independence estimates as non-informative prior which get adjusted by the data through the posterior mean.

Cell ID	111	112	121	122	211	212	221	222
Actual	.1188	.1749	.0924	.1287	.1155	.1882	.0957	.0858
Independent	.1205	.1732	.0907	.1304	.1322	.1714	.0790	.1025
Our	.1119	.1219	.1856	.1778	.0904	.0896	.1178	.0963
Proposal	(.0277)	(.0270)	(.0275)	(.0269)	(.0292)	(.0285)	(.0288)	(.0295)

Table 8: Summary for the Results for Survey Data

# 5 Summary

This paper has put forward a technique for simplifying Bayesian calculations for multiway contingency tables for estimating the unknown cell proportions for given marginals. Knowledge of prior cell-proportions in the form of a Dirichlet distribution is assumed. However, in the absence of such prior, it is illustrated that the probabilities obtained using the independence assumption of the factors may be used as non-informative prior. The resulting estimates are often quite close to the actual observed cell proportions. The technique is illustrated using some real data.

# Acknowledgments

This research is partially supported by a research grant from Bell University Labs to Y. P. Chaubey and F. Nebebe.

## References

- [1] Bishop, Y. M. M., Feinberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice, Cambridge, MA: MIT Press.*
- [2] Putler, D. S., Kalyanam, K. and Hodges, J. S. (1996). A Bayesian Approach for Estimating Target Market Potential with Limited Geodemographic Information. *Journal of* Market Research, **33**, 134-149.