Computational Statistics & Data Analysis 7 (1989) 245-258
North-Holland

Model-based Bayesian methods for estimating cell proportions in cross-classification tables having ordered categories

Alan AGRESTI

Department of Statistics, University of Florida, Gainesville, FL 32611, USA

Christy CHUANG

Biostatistics Unit, Upjohn Company, Kalamazoo, MI 49001, USA

Received 27 February 1987
Revised 8 June 1988

Abstract: Bayesian methods are suggested for estimating proportions in the cells of cross-classification tables having at least one classification with ordered categories. These methods utilize models for cell proportions that incorporate the category orderings. The resulting estimators are smoother and can be much more efficient than the sample proportions, yet they are consistent even if the model chosen for the smoothing does not hold. Two approaches are considered: (1) Bayes estimators using a Dirichlet prior distribution for the proportions; (2) Bayes estimators based on normal prior distributions for association parameters in the saturated loglinear model. In each case, the means of the prior distributions are chosen to satisfy a model for ordered categorical data, such as the uniform association model. Empirical Bayes versions of the two analyses are also given.

Keywords: Dirichlet prior, EM algorithm, Empirical Bayes, Loglinear models, Ordinal variables, Smoothing methods, Uniform association model.

1. Introduction

In recent years, much attention has been devoted to the development of models for analyzing cross-classification tables in which classifications have ordered categories. In some applications, however, it is important to estimate cell proportions in the table, yet there is no reason to expect a certain model to describe the data well. Model-based estimators are inconsistent, when the model does not hold. On the other hand, the sample proportions may not be desirable estimators, especially if the data are sparse. In this article we suggest two Bayesian approaches to smoothing the sample proportions. In these approaches, a model still provides part of the mechanism for smoothing the data, in the sense that the methods produce a shrinkage of the sample proportions toward a set of proportions satisfying the model. The estimators combine good characteristics of sample proportions and of estimators that are completely model-based. Like sample proportions (and unlike model-based estimators), they are consistent. Like model-based estimators (and unlike sample proportions), the estimators incorporate the ordinal nature of the data, giving smoother values that can have much smaller total mean square error than the sample proportions.

In Section 2 we suggest a Bayesian analysis that applies directly to the cell proportions. We use a variation of the Fienberg and Holland (1970, 1973) approach of giving a Dirichlet prior distribution to the cell proportions. Our Dirichlet prior distribution has expected value components satisfying a simple model for ordinal data, such as Goodman's (1979) uniform association model. In a corresponding empirical Bayes approach, the expected value components in the prior distribution are obtained from the regular maximum likelihood (ML) fit of the ordinal model. In Section 3 we give a Bayesian analysis that applies to the parameters of the saturated loglinear model, rather than directly to the cell proportions. The expected values of normal prior distributions for the association parameters in the saturated model again follow the structure of a simple model for ordinal data. An empirical Bayes method that uses the EM algorithm is given for estimating the parameters in the prior distributions.

Section 4 gives an example, and Section 5 gives results of a Monte Carlo study in which empirical Bayes estimators based on Dirichlet prior distributions are compared to the sample proportion estimator. The smoothed estimators are seen to be much more efficient than the sample proportion, their relative advantage increasing as the number of cells increases, as the sample size decreases, or as the model utilized for the smoothing better approximates the pattern for the true proportions.

Most of our ideas are presented in the context of estimating proportions in the cells of a two-way table in which at least one of the classifications is ordered. We assume that the sample cell counts $\boldsymbol{n}' = (n_{11}, \ldots, n_{rc})$ in the r-by-c cross-classification of X and Y have a multinomial $(n, \{\pi_{ij}\})$ distribution, where $n = \sum\sum n_{ij}$. Denote the expected values of the $\{n_{ij}\}$ by $\{m_{ij}\}$. The local odds ratios

$$\theta_{ij} = m_{ij} m_{i+1,j+1} / m_{i,j+1} m_{i+1,j}, \quad 1 \leqslant i \leqslant r-1,\ 1 \leqslant j \leqslant c-1,$$

are useful for describing properties of models for the association between X and Y.

Several simple and useful models are special cases of the model

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \beta \mu_i \nu_j \tag{1.1}$$

considered by Goodman (1979). For this model,

$$\log \theta_{ij} = \beta(\mu_{i+1} - \mu_i)(\nu_{j+1} - \nu_j).$$

For the special case $\{\mu_i = i\}$ and $\{\nu_j = j\}$ this is called the *uniform association* (U) model, since $\{\log \theta_{ij} = \beta\}$. When the $\{\mu_i\}$ are unspecified parameters and the $\{\nu_j\}$ are monotone scores, this is called the *row-effects* model, and when the $\{\nu_j\}$ are unspecified and the $\{\mu_i\}$ are monotone scores it is called the *column effects* model. One reason for the importance of this structural form is that it contains as a special case a discrete version of the bivariate normal distribution (Goodman 1985). This model is useful for both of the smoothing methods described in this article.

2. Using an ordinal Dirichlet prior distribution for cells proportions

Fienberg and Holland (1970, 1973) described a simple Bayesian approach in which the unknown cell proportions $\boldsymbol{\pi}' = (\pi_{11}, \dots, \pi_{rc})$ have a Dirichlet prior distribution with parameter $\boldsymbol{\beta}' = (\beta_{11}, \dots, \beta_{rc})$; that is, the prior density function is

$$f(\boldsymbol{\pi}) = \frac{\Gamma(\sum\sum\beta_{ij})}{\prod_i\prod_j \Gamma(\beta_{ij})} \prod_i\prod_j (\pi_{ij})^{\beta_{ij}-1},$$

$$0 \leqslant \pi_{ij} \leqslant 1 \text{ for all } i \text{ and } j, \ \sum\sum\pi_{ij} = 1,$$

where all $\beta_{ij} > 0$. The prior mean of π_{ij} is $\gamma_{ij} = \beta_{ij}/K$, where $K = \Sigma\Sigma\beta_{ij}$. The posterior distribution of $\boldsymbol{\pi}$ is also Dirichlet, with parameter $\boldsymbol{\beta}^* = \boldsymbol{\beta} + \boldsymbol{n}$. Let $\alpha = K/(n+K)$. The Bayes estimator for squared error loss is the posterior mean,

$$E(\pi_{ij} \mid \boldsymbol{n}) = (1-\alpha)p_{ij} + \alpha\gamma_{ij} \tag{2.1}$$

which is a weighted average of the sample proportion $p_{ij} = n_{ij}/n$ and the prior mean γ_{ij}.

Formula (2.1) suggests that K can be interpreted as the number of observations that the prior information represents. The value of K for which the total mean squared error (MSE) is minimized is

$$K(\boldsymbol{\pi}, \boldsymbol{\gamma}) = \left[1 - \sum\sum\pi_{ij}^2\right] \Big/ \left[\sum\sum(\gamma_{ij} - \pi_{ij})^2\right]. \tag{2.2}$$

Fienberg and Holland suggested an estimator of π_{ij} having form (2.1), but with K replaced by the ML estimator $K(\boldsymbol{p}, \boldsymbol{\gamma})$ of $K(\boldsymbol{\pi}, \boldsymbol{\gamma})$. For this choice,

$$\alpha = \left[1 - \sum\sum p_{ij}^2\right] \Big/ \left[n\sum\sum(\gamma_{ij} - p_{ij})^2 + \left(1 - \sum\sum p_{ij}^2\right)\right]. \tag{2.3}$$

Following the arguments in Brown and Rundell (1985) for kernel estimates, one could instead obtain an unbiased estimator of the total mean square error, and then solve for the value of α that minimizes that estimator. This gives the "minimum unbiased risk estimator", for which

$$\alpha = \left[1 - \sum\sum p_{ij}^2\right] \Big/ \left[(n-1)\sum\sum(\gamma_{ij} - p_{ij})^2\right]$$

and there is greater smoothing than with the Fienberg-Holland estimator. Alternative estimators could be used for K, such as those discussed by Bishop et al.

(1975), pp. 430–432) and such as the ratio unbiased and two-step estimators suggested by Ighadaro and Santner (1982).

For ordinal variables, we suggest giving the $\{\gamma_{ij}\}$ a pattern that reflects trends expected in the association. Unless one wishes to posit specific values for the $\{\gamma_{ij}\}$, it is probably easiest to select values that satisfy a simple ordinal model. For instance, if both variables are ordinal, one often expects (at least approximately) a monotonic association of the type in which the $\{\log\theta_{ij}\}$ are uniformly of one sign. Then it is natural to let the $\{\gamma_{ij}\}$ satisfy model (1.1) with monotone scores. For the U model, for instance, the choices of the common local log odds ratio β and the row and column marginal probabilities determine the $\{\gamma_{ij}\}$.

One can bypass having to choose the $\{\gamma_{ij}\}$ by using an empirical Bayes approach, for which the $\{\gamma_{ij}\}$ depend on the data. Fienberg and Holland suggested $\{\hat{\gamma}_{ij}=p_{i+}p_{+j}\}$, where $p_{i+}=\Sigma_j p_{ij}$ and $p_{+j}=\Sigma_i p_{ij}$, which shrinks the sample proportions towards the fit of the independence model. To utilize the category orderings, we instead recommend using $\{\hat{\gamma}_{ij}\}$ that are ML fitted probabilities for a simple ordinal model. If we select the U model, for instance, then the component means of the fitted Dirichlet prior distribution for $\{\pi_{ij}\}$ match the data in the marginal distributions and in the correlation, since the likelihood equations for the U model are $\{\hat{\gamma}_{i+}=p_{i+}\}$, $\{\hat{\gamma}_{+j}=p_{+j}\}$, and $\Sigma\Sigma ij\hat{\gamma}_{ij}=\Sigma\Sigma ijp_{ij}$. The resulting posterior estimator has the appealing property of being a weighted average of the sample proportion and the ML fitted proportion for the U model. For fixed n, the weight given to the sample proportion decreases as the fit of the U model improves.

This strategy can be suitably modified for cross classifications of ordinal with nominal variables. For instance, if the row variable is nominal and the column variable is ordinal, one often expects (at least approximately) the conditional distributions within the rows to be stochastically ordered on the ordinal variable. Then, it is reasonable to let the $\{\hat{\gamma}_{ij}\}$ be ML estimates of cell proportions for a model that has this property, such as the row effects model. For that choice, the component expected values of the fitted Dirichlet prior distribution is an rc-vector that matches the sample proportions in the marginal distributions and in the row means $\{\Sigma_j v_j p_{ij}/p_{i+},\ i=1,\dots,r\}$.

3. Using ordinal normal prior distributions for loglinear model parameters

Leonard (1975) and Laird (1978) proposed estimating proportions in a two-way table through a Bayesian analysis of the parameters of the saturated loglinear model,

$$\log m_{ij}=\mu+\lambda_i^X+\lambda_j^Y+\lambda_{ij}.$$

Laird let the marginal parameters $\{\lambda_i^X\}$ and $\{\lambda_j^Y\}$ have independent uniform (improper) prior distributions over the real line, subject to the constraints $\Sigma\lambda_i^X=\Sigma\lambda_j^Y=0$, and she let the association parameters $\{\lambda_{ij}\}$ have independent $N(0,\sigma^2)$ distributions. She also suggested empirical Bayes analyses in which one estimates σ^2 by finding the value that maximizes an approximation for the marginal distribution of $\boldsymbol{n}$, evaluated at the observed data.

The use of zero for the prior means for the $\{\lambda_{ij}\}$ provides a shrinkage towards the independence model. For ordinal data, one could instead let the prior means be terms in a simple structural model that describes ordinal relationships. For instance, we suggest taking $E(\lambda_{ij}) = \beta\mu_i\nu_j$, corresponding to model (1.1). Our analysis is an adaptation of the one presented by Laird, in which we provide this additional prior structure. Suppose both variables are ordinal but there is no natural set of category scores. If one expects some (unspecified) monotonic form of association, it is simplest to use the scoring $\{\mu_i = i-(r+1)/2\}$ and $\{\nu_j = j-(c+1)/2\}$, for which the mean of the prior distribution for λ_{ij} is the association term in the U model. Specification of β in this prior mean reflects beliefs about the direction and strength of association. Specification of σ reflects beliefs about the degree to which the $\{\lambda_{ij}\}$ approximate a uniform association structure with that strength of association.

Let $\boldsymbol{\theta} = (\lambda_1^X,\ldots,\lambda_{r-1}^X, \lambda_1^Y,\ldots,\lambda_{c-1}^Y, \lambda_{11},\ldots,\lambda_{rc})$. The posterior density function of $\boldsymbol{\theta}$ has the form

$$h(\boldsymbol{\theta}\,|\,\boldsymbol{n};\ \beta,\ \sigma) \propto \exp\left\{\sum\sum n_{ij}\log(m_{ij}) - \sum\sum(\lambda_{ij}-\beta\mu_i\nu_j)^2/2\sigma^2\right\}.$$

We take the approach of Leonard (1975) and Laird (1978) of using the mode $\boldsymbol{\theta}^*$ of the posterior distribution as the estimator of $\boldsymbol{\theta}$. This estimator corresponds to an estimator $\boldsymbol{\pi}^*$ of $\boldsymbol{\pi}$, for which the expected frequency estimates $\{m_{ij}^* = n\pi_{ij}^*\}$ satisfy the equations

$$m_{i+}^* = n_{i+}, \quad i=1,\ldots,r,$$

$$m_{+j}^* = n_{+j}, \quad j=1,\ldots,c,$$

$$m_{ij}^* = n_{ij} - (\lambda_{ij}^* - \beta\mu_i\nu_j)/\sigma^2, \quad i=1,\ldots,r,\ j=1,\ldots,c.$$

From the second derivative matrix of the log posterior with respect to $\boldsymbol{\theta}$, it follows that the $\{m_{ij}^*\}$ satisfying these equations are unique. If all $n_{i+}>0$ and all $n_{+j}>0$, then all $\pi_{ij}^*>0$, even if some $n_{ij}=0$. The equations imply that as $\sigma^2\to\infty$, $\{m_{ij}^*\to n_{ij}\}$, whereas as as $\sigma^2\to 0$, $\lambda_{ij}^*\to\beta\mu_i\nu_j$ and hence the $\{m_{ij}^*\}$ converge to values that satisfy model (1.1) with association term identical to the mean of the prior distribution. For fixed $\sigma^2>0$, as $n\to\infty$, $\pi_{ij}^* = p_{ij} + O_p(1/n)$.

To avoid having to choose β and σ^2 in the prior distributions for $\{\lambda_{ij}\}$, one can instead use an empirical Bayes solution. Here, we adapt the approach suggested by Laird (1978) and also by Chuang (1982) for a different model. The posterior density is related to the prior density $g(\boldsymbol{\theta})$ and the likelihood $f(\boldsymbol{n}\,|\,\boldsymbol{\theta})$ by

$$h(\boldsymbol{\theta}\,|\,\boldsymbol{n};\ \beta,\ \sigma) = g(\boldsymbol{\theta})f(\boldsymbol{n}\,|\,\boldsymbol{\theta})/m(\boldsymbol{n};\ \beta,\ \sigma),$$

where $m(\boldsymbol{n};\ \beta,\ \sigma)$ denotes the marginal probability function of $\boldsymbol{n}$. Viewing the marginal distribution as a function of $(\beta,\ \sigma)$, for a given $\boldsymbol{n}$, we estimate the parameters in the prior distribution by maximizing this "marginal likelihood".

Using the same arguments given by Laird, we obtain equations for maximizing $m(\boldsymbol{n};\ \beta,\ \sigma)$,

$$E\left(\sum\sum\lambda_{ij}^2 \mid \boldsymbol{n},\ \beta,\ \sigma\right) = rc\sigma^2 + \beta^2 \sum\sum \mu_i^2 \nu_j^2 \tag{3.1}$$

and

$$E\left(\sum\sum\mu_i\nu_j\lambda_{ij} \mid \boldsymbol{n},\ \beta,\ \sigma\right) = \beta \sum\sum \mu_i^2 \nu_j^2 . \tag{3.2}$$

The EM algorithm can be applied to solve equations (3.1) and (3.2) as follows. Given current values $\sigma^{(p)}$ and $\beta^{(p)}$, for the E step one calculates

$$t^{(p)} = E\left(\sum\sum\lambda_{ij}^2 \middle| \boldsymbol{n},\ \beta^{(p)},\ \sigma^{(p)}\right)$$

and

$$u^{(p)} = E\left(\sum\sum\mu_i\nu_j\lambda_{ij} \middle| \boldsymbol{n},\ \beta^{(p)},\ \sigma^{(p)}\right).$$

Then for the M step one lets

$$\beta^{(p+1)} = u^{(p)}\left(\sum\sum\mu_i^2\nu_j^2\right)^{-1}$$

and

$$\left(\sigma^{(p+1)}\right)^2 = \left[t^{(p)} - \left(\beta^{(p+1)}\right)^2 \sum\sum\mu_i^2\nu_j^2\right] / rc.$$

The conditional expectations require integration with respect to the posterior distribution. One can approximate this integral by replacing the posterior distribution by the normal distribution $N(\boldsymbol{\theta}^*, \Sigma^*)$ having the same mode and whose log has the same curvature at the mode as the log of the true posterior. Thus, the E step of the EM algorithm is implemented by taking

$$t^{(p)} = \sum\sum\left(\lambda_{ij}^*\right)^2 + \sum\sigma_k^* \quad \text{and} \quad u^{(p)} = \sum\sum\mu_i\nu_j\lambda_{ij}^*$$

where $\{\sigma_k^*\}$ are the estimated variances, obtained from Σ^*, of the rc values $\{\lambda_{ij}^*\}$ obtained in the p^{th} iteration. The matrix Σ^* has the same form as given by Laird (1978, p. 586).

Since

$$\left[\beta^{(p+1)}\right]^2 \sum\sum\mu_i^2 v_j^2 = \frac{\left[u^{(p)}\right]^2}{\sum\mu_i^2\nu_j^2} = \frac{\left[\sum\mu_i\nu_j\lambda_{ij}^*\right]^2}{\sum\mu_i^2\nu_j^2} \leqslant \sum\sum\left(\lambda_{ij}^*\right)^2$$

it follows that $[\sigma^{(p+1)}]^2 \geqslant (\sum\sigma_k^*)/rc$. Hence, the empirical Bayes estimate of σ^2 is nonnegative.

For this application the EM algorithm may converge extremely slowly, and there need not be a unique solution. In fact, ($\sigma = 0$, arbitrary β) are always roots of these equations. We suggest using a wide variety of initial values for β and σ in order to check whether the obtained solution depends on the initial choice. To guard against choosing an inappropriate solution, we also suggest comparing results to those obtained with an alternative approximation. For instance, we adapted an alternative approach suggested by Laird. Since

$$m(\boldsymbol{n};\, \beta, \sigma) = g(\boldsymbol{\theta};\, \beta, \sigma) f(\boldsymbol{n} \mid \boldsymbol{\theta}) / h(\boldsymbol{\theta} \mid \boldsymbol{n};\, \beta, \sigma) \quad \text{for all } \boldsymbol{\theta}, \tag{3.3}$$

it can be approximated by dividing the product of the likelihood and prior density by the normal approximation discussed previously for the posterior density. The resulting approximation for the marginal distribution should be reasonable when the normal approximation for the posterior distribution is evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. Using this substitution for each $\boldsymbol{\theta}$ term in (3.3), we calculated numerically the approximate marginal distribution for a rectangular grid of (β, σ) values, in order to determine an approximation for the (β, σ) value that maximizes the marginal distribution.

For several examples studied by the authors, both approaches for determining (β, σ) gave very similar results. Because convergence was slow with the EM algorithm, however, and since there exist multiple solutions to equations (3.1) and (3.2), we have found the second approximation more useful, particularly when $\sigma = 0$. In this approach, a very broad range of (β, σ) values was first used to determine the general behavior of the marginal distribution and to determine the region in which its maximum occurs, and then a refined grid of (β, σ) values was used to better determine the location of the maximum. Computer programs for obtaining the empirical Bayes solution using these approximations are available upon request from Dr. Chuang.

An interesting characteristic of the empirical Bayes approach is that if the standard ML fit of model (1.1) is exceptionally good, then the marginal distribution of $\boldsymbol{n}$ is sometimes maximized at a (β, σ) pair for which $\sigma = 0$. When this happens, we conjecture that the β value in that maximizing pair is identical to the standard ML estimate $\hat{\beta}$ for model (1.1). Our reasoning is as follows. Consider expression (3.3) for fixed $\boldsymbol{n}$ and fixed β. This expression holds for all $\boldsymbol{\theta}$, and hence it applies at the posterior mode $\boldsymbol{\theta}^*$. Now as $\sigma \downarrow 0$ the posterior density of $\boldsymbol{\theta}$ loses its dependence on $\boldsymbol{n}$, becoming more similar to the prior density, and it seems as if the ratio of the prior and posterior densities evaluated at the posterior mode would converge to 1, hence losing its dependence on β. Thus, as $\sigma \downarrow 0$, finding the β value that maximizes $m(\boldsymbol{n};\, \beta, \sigma)$ becomes more similar to the problem of maximizing the regular likelihood $f(\boldsymbol{n} \mid \boldsymbol{\theta})$ subject to the constraints $\{\lambda_{ij} = \beta\mu_i\nu_j\}$ for $\boldsymbol{\theta}$ implied by $\sigma = 0$. This latter maximization corresponds to ML estimation of β in model (1.1). If our conjecture is true, then the posterior estimate of λ_{ij} is $\hat{\beta}\mu_i\nu_j$, and since $m^*_{i+} = n_{i+}$ and $m^*_{+j} = n_{+j}$ for all i and j, it follows that the $\{m^*_{ij}\}$ are identical to the ML estimates for model (1.1). This result gives an interesting interpretation to the empirical Bayes solution: If model (1.1) fits extremely well, one estimates the cell proportions using the ML expected frequency estimates based directly on that model; otherwise, one uses estimates that correspond to a less severe smoothing of the sample proportions.

When the row variable is nominal and the column variable is ordinal, one could still use the $N(\beta\mu_i\nu_j, \sigma^2)$ prior distribution for λ_{ij}, but it no longer is appropriate to use monotone scores for $\{\mu_i\}$. One could instead use the parameterization $N(\tau_i\nu_j, \sigma^2)$, where $\{\nu_j\}$ are fixed (e.g., equal-interval) and σ^2 and the $\{\tau_i\}$, satisfying $\Sigma\tau_i = 0$, are parameters. Then, the mean of the prior distribution is the association term in the row effects model. The empirical Bayes approaches discussed previously can be adapted to this situation.

Like the approach of Section 2, the Bayesian procedure discussed in this section does not smooth the row or column marginal proportions. Such a smoothing can be obtained, however, by using proper prior distributions for the marginal parameters. For instance, one could let the $\{\lambda_j^X\}$ be independent $N(0, \sigma_1^2)$ and the $\{\lambda_j^Y\}$ be independent $N(0, \sigma_2^2)$.

4. Example

Next we illustrate the smoothing approaches by estimating cell proportions for Table 1, taken from Maxwell (1961), giving data on severity of disturbed dreams

Table 1

Expected frequency estimates based on (1) sample proportions, (2) ML estimates for model (1.1), (3) Empirical Bayes estimates with Dirichlet prior distribution for $\{\pi_{ij}\}$, (4) Empirical Bayes estimates with normal prior distributions for $\{\lambda_{ij}\}$

Age	Estimate	Severity of Disturbances of Dreams			
		Not Severe			Very Severe
		1	2	3	4
5-7	1	7	4	3	7
	2	4.80	3.39	5.23	7.58
	3	5.78	3.66	4.24	7.32
	4	5.34	3.51	4.83	7.31
8-9	1	10	15	11	13
	2	16.41	9.09	11.01	12.50
	3	13.56	11.72	11.01	12.72
	4	14.61	10.74	10.98	12.66
10-11	1	23	9	11	7
	2	21.41	9.76	9.73	9.10
	3	22.12	9.42	10.29	8.17
	4	22.09	9.51	9.87	8.54
12-13	1	28	9	12	10
	2	30.72	11.53	9.46	7.29
	3	29.51	10.41	10.59	8.49
	4	29.67	11.02	10.22	8.09
14-15	1	32	5	4	3
	2	26.67	8.24	5.57	3.53
	3	29.04	6.80	4.87	3.29
	4	28.29	7.22	5.09	3.41

Source of data: Maxwell (1961).

for a sample of 223 boys. In utilizing model (1.1), we assigned midpoint scores $\mu' = (6, 8.5, 10.5, 12.5, 14.5)$ to the age variable and equal-interval scores $\{ \nu_j = j - 3.5)$ to severity of disturbed dreams. Table 1 contains the estimated expected frequency estimates based on (1) the sample proportions (i.e., these are the cell counts), (2) the ML fit of model (1.1), (3) the empirical Bayes approach with a Dirichlet prior distribution for cell proportions that smooths towards the fit of model (1.1) (namely, (2.1) with α obtained using (2.3) with γ_{ij} replaced by the fitted values for that model), and (4) the empirical Bayes approach using $\{ N(\beta\mu_i\nu_j, \sigma^2)\}$ prior distributions for $\{\lambda_{ij}\}$ in the saturated loglinear model.

Model (1.1) with the indicated scores fits these data quite well. The likelihood-ratio goodness-of-fit statistic equals 14.6, based on 11 degrees of freedom, and the ML estimate of β equals -0.097.

When we use the Dirichlet prior distribution for $\{\pi_{ij}\}$ with $\{\hat{\gamma}_{ij}\}$ that are the ML fitted probabilities for the model (1.1), then the weight given to the model (1.1) estimates is $\alpha = 0.56$. The moderately strong weight reflects the good fit obtained with that model.

For the empirical Bayes approach with prior distributions for the $\{\lambda_{ij}\}$, the EM algorithm produces $\beta = -0.092$ and $\sigma = 0.184$ for the parameterization of the normal distributions. Direct approximation of the marginal distribution of $\boldsymbol{n}$ suggested that its value for the observed data is maximized approximately when $\beta = -0.094$ and $\sigma = 0.187$. The relatively small value for σ (compared to the values of $\beta\mu_i\nu_j$ in the corners of the table, for instance) again reflects the good fit obtained with the model (1.1). The values $\beta = -0.092$ and $\sigma = 0.184$ were used in

Table 2
Expected Frequency Estimates based on (1) Sample Proportions, (2) ML Estimates for U Model, and Empirical Bayes Estimates with Normal Prior Distributions for $\{\lambda_{ij}\}$, and (3) Empirical Bayes Estimates with Dirichlet Prior Distribution for $\{\pi_{ij}\}$

Mental Health Status	Estimate	Parents' Socioeconomic Status					
		A	B	C	D	E	F
Well	1	64	57	57	72	36	21
	2	65.3	54.2	55.9	65.3	39.0	27.3
	3	64.9	55.0	56.2	67.1	38.1	25.6
Mild	1	94	94	105	141	97	71
Symptom	2	104.4	94.9	107.2	137.0	89.6	68.8
Formation	3	101.5	94.7	106.6	138.1	91.6	69.4
Moderate	1	58	54	65	77	54	54
Symptom	2	50.2	49.9	61.7	86.4	61.8	52.0
Formation	3	52.3	51.1	62.6	83.8	59.6	52.6
Impaired	1	46	40	60	94	78	71
	2	42.1	45.9	62.2	95.3	74.7	68.8
	3	43.2	44.3	61.6	94.9	75.6	69.4

Source of data: Srole et al. (1962)

the normal prior distributions that generated the posterior distribution corresponding to the expected frequency estimates given in Table 1. For these data, the Dirichlet and normal-based priors gave very similar results in terms of the smoothings provided of the sample cell counts.

Table 2 is an example of data for which the empirical Bayes approach for parameters in the saturated loglinear model gives a degenerate prior distribution and hence gives cell proportion estimates identical to the ML estimates for model (1.1). The data are taken from Srole et al. (1978, p. 289), and were analyzed in Goodman (1979). Theory and research in mental health studies have suggested that mental health improves with increasing socioeconomic status (SES) of subjects or their parents (see, e.g., Dohrenwend and Dohrenwend, 1969). Hence, it makes sense to smooth towards a model, such as the U model, that represents a monotonic association. In fact, the U model fits the data extremely well, with likelihood-ratio goodness-of-fit statistic equal to 9.9, based on df = 14.

When we use the Dirichlet prior distribution for $\{\pi_{ij}\}$ with $\{\hat{\gamma}_{ij}\}$ that are ML fitted probabilities for the U model, then $\alpha = 0.72$, so the smoothed estimates are quite similar to the ML estimates for the U model. For the empirical Bayes approach with $N(\beta\mu_i\nu_j, \sigma^2)$ prior distributions for the $\{\lambda_{ij}\}$ in the saturated loglinear model, the EM algorithm and direct approximation of the marginal distribution of $\boldsymbol{n}$ given (β, σ) both suggested the use of $\beta = 0.09$ and $\sigma = 0.00$ in the prior distributions. In fact, $\hat{\beta} = 0.091$ is the ML estimate of β for the U model.

5. Improvement in estimation using smoothed estimators

In this section we give the results of a Monte Carlo study that illustrates how smoothed Bayes estimators can be substantially better than sample proportions. This study used only the empirical Bayes estimator for the Dirichlet prior approach of Section 2, because computation time is much less for the normal-prior based estimator of Section 3.

The study was also designed to study how various factors affect the behavior of the smoothed vs. unsmoothed estimators. We compared the mean square error of the sample proportions and of the smoothed estimators, by varying

1. Table size
   a. 3×3
   b. 6×6
2. Sample size
   a. $n = 50$
   b. $n = 200$
3. Whether the U model holds
   a. Yes
   b. No – Model (1.1) holds with scores $\{1, 2.5, 3\}$ for 3×3 case, $\{1, 2.8, 4.2, 5.2, 5.8, 6\}$ for 6×6 case.

Table 3
Mean Square Errors for Estimators of Cell Proportions

$r \times c$	β	n	Estimator			
			Sample Proportion	Uniform-smoothed	Independence-smoothed	Uniform model
U Model Holds						
3×3	0.1	50	0.099	0.067	0.060	0.062
		200	0.099	0.068	0.064	0.062
	0.4	50	0.098	0.069	0.076	0.064
		200	0.098	0.067	0.095	0.062
6×6	0.1	50	0.027	0.012	0.013	0.009
		200	0.027	0.012	0.016	0.009
	0.4	50	0.027	0.014	0.018	0.012
		200	0.027	0.014	0.024	0.012
U Model Does Not Hold						
3×3	0.1	50	0.099	0.067	0.060	0.063
		200	0.099	0.068	0.065	0.064
	0.4	50	0.098	0.072	0.078	0.071
		200	0.098	0.084	0.098	0.097
6×6	0.1	50	0.027	0.013	0.013	0.010
		200	0.027	0.013	0.017	0.012
	0.4	50	0.027	0.016	0.019	0.016
		200	0.027	0.020	0.024	0.032

4. Strength of association
   a. $\beta = 0.1$ in model (1.1)
   b. $\beta = 0.4$ in model (1.1)

Uniform marginal probabilities $\{\pi_{i+} = 1/r\}$ and $\{\pi_{+j} = 1/c\}$ were used for all cases. The Dirichlet prior distribution was based on the ML fit of the U model, namely (2.1) where α was calculated using (2.3) with γ_{ij} replaced by the ML estimate of π_{ij} using the U model. For illustrative purposes, we also compared this estimator and the sample proportion to the ML estimator based completely on the U model and to the Bayes estimator whose Dirichlet prior distribution is based on the ML fit of the independence model (i.e., (2.1) where α is calculated using (2.3) with γ_{ij} replaced by $p_{i+}p_{+j}$).

For each estimator $\hat{\pi}_{ij}$ of π_{ij}, Table 3 contains the value of

$$\frac{n\sum_{a=1}^{M}\sum_i\sum_j\left(\hat{\pi}_{ij,a} - \pi_{ij}\right)^2}{Mrc} \tag{5.1}$$

for each of the sixteen combinations of conditions, where $\hat{\pi}_{ij,a}$ denotes the value of $\hat{\pi}_{ij}$ in the a^{th} randomly generated table. The $M = 5000$ simulations used the GGMTN multinomial generator in IMSL, on an IBM 3081 computer. For this M value, the standard errors of the values in (5.1) are all less than 0.001. For the sample proportion estimator, we report the exact expected value of (5.1), which is $(1 - \Sigma\Sigma\pi_{ij}^2)/rc$.

When the U model holds, the amount of smoothing is substantial for the U-smoothed empirical Bayes estimator. Its mean square error is considerably smaller than that for the sample proportion, and is nearly as small as that for the estimator based completely on the U model. Not surprisingly, the U-smoothed estimator improves on the independence-smoothed estimator as the strength of association β increases, and as the sample size n increases.

Similar remarks apply when model (1.1) holds but the U model does not, though as n increases it naturally becomes more advantageous to use the smoothed proportion estimators instead of those based completely on the U model. However, it is only for the largest values of $r \times c$, β, and n considered (the last row in Table 3) that the inconsistency of the U model estimator starts to affect its performance seriously. The uniform-smoothed estimator does well in all cases considered. The MSE is quite a bit smaller than that for the sample proportion even when the U model does not hold and with n as large as 200.

6. Generalizations and alternative approaches

This article has considered only two-way tables. In principle, the arguments extend to multi-dimensional tables having some ordinal classifications. For the approach of Section 2, one selects an appropriate ordinal model for the mean of the Dirichlet prior to satisfy. For many purposes it would be adequate to choose a model that structures the two-factor associations and excludes three-factor or higher interactions. Examples of such models were given by Clogg (1982) and Agresti and Kezouh (1983). For the approach of Section 3, one lets the highest-order interaction terms in the saturated model have independent normal prior distributions, where the means in the prior distributions are terms in an ordinal model that one expects to approximate the true form of the interaction. The Dirichlet approach is considerably simpler to implement particularly for tables of several dimensions.

There are many ways other than Bayesian methods to smooth ordered categorical data. Kernel methods were discussed by Titterington (1980), Titterington and Bowman (1985), and Brown and Rundell (1985). An advantage of these, compared to most of the Bayes methods discussed here, is that the marginal proportions also are smoothed. A disadvantage is that the choice of kernel method is usually ad hoc rather than theory-based. In some of our simulations we used the kernel estimator $\pi^* = (I + \alpha G)'p$, where the matrix G has all $g_{ii} = -1$ and gives influence of p_{ab} on π^*_{cd} proportional to

$$0.5^{(a-c)^2+(b-d)^2}$$

and where α was chosen to minimize an unbiased estimate of the total mean square error. Compared to the empirical Bayes approach of Section 2, the kernel approach was more successful (i) as the true cell proportions were more nearly constant, (ii) as the model used for the Bayesian smoothing fitted the true proportions more poorly, and (iii) as the sample size decreased.

Simonoff (1983) and Titterington and Bowman (1985) discussed the penalized likelihood approach, in which the smoothed estimator is obtained by maximizing

$$L = \text{log likelihood} - \Phi(\pi),$$

where Φ is a roughness penalty; that is, Φ is a function that decreases as π is more smooth, in some sense. For two-way tables Simonoff (1983) suggested the penalty function $\Phi(\pi) = \eta\Sigma\Sigma(\log\ \theta_{ij})^2$ involving the local odds ratios $\{\theta_{ij}\}$, which has the effect of shrinkage towards the independence estimator. For ordinal data, one could select a function that penalizes for departures from smoothness given by a certain type of ordinal model. For instance, one could let

$$\Phi(\pi) = \eta\sum\sum(\log\ \theta_{ij} - \beta)^2,$$

which has the effect of penalizing more when the estimates move farther from a uniform association fit. This approach, a kernel approach, or a Bayes or empirical Bayes approach generally produce estimators that, for sparse ordinal data, are much preferable to the sample proportions.

Acknowledgments

We thank Ming Yang for computing assistance. Dr. Agresti's research was partially supported by grant GM 33210 from the National Institutes of Health.

References

A. Agresti and A. Kezouh, Association models for multidimensional cross-classifications of ordinal variables, *Communications in Statistics* **A12** (1983) 1261–1276.

Y.M.M. Bishop, S.E. Fienberg and P.W. Holland, *Discrete Multivariate Analysis* (MIT Press, Cambridge, 1975).

P.J. Brown and P.W.K. Rundell, Kernel estimates for categorical data, *Technometrics* **27** (1985) 293–299.

C. Chuang, Empirical Bayes methods for a two-way multiplicative-interaction model, *Communications in Statistics* **A11** (1982) 2977–2989.

C. Clogg, Some models for the analysis of association in multiway cross-classifications having ordered categories, *J. Amer. Statist. Assoc.* **77** (1982) 803–815.

B.P. Dohrenwend and B.S. Dohrenwend, *Social Status and Psychological Disorder: A Causal Inquiry* (Wiley, New York, 1969).

S.E. Fienberg and P.W. Holland, Methods for eliminating zero counts in contingency tables, in: G.P. Patil, (Ed.) *Random Counts on Models and Structures* (Pennsylvania State University Press, University Park, 1970).

S.E. Fienberg and P.W. Holland, Simultaneous estimation of multinomial cell probabilities, *J. Amer. Statist. Assoc.* **68** (1973) 683–691.

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

L.A. Goodman, The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries, *Ann. Statist.* **13** (1985) 10–69.

A. Ighodaro and T. Santner, Ridge type estimators of multinomial cell probabilities, in: S. Gupta and J.O. Berger (Eds.) *Statistical Decision Theory and Related Topics III*, Vol. 2, (Academic Press, New York, 1982).

N. Laird, Empirical Bayes methods for two-way contingency tables, *Biometrika* **65** (1978) 581–590.

T. Leonard, Bayesian estimation methods for two-way contingency tables, *J. Roy. Statist. Soc.* **B37** (1975) 23–37.

A.E. Maxwell, *Analysing Qualitative Data* (Methuen, London, 1961).

J.S. Simonoff, A penalty function approach to smoothing large sparse contingency tables, *Ann. Statist.* **11** (1983) 208–218.

L. Srole, T.S. Langner, S.T. Michael, M.K. Opler and T.A.C. Rennie, *Mental Health in the Metropolis: The Midtown Manhattan Study* (McGraw-Hill, New York, 1962).

D.M. Titterington, A comparative study of kernel-based density estimates for categorical data, *Technometrics* **22** (1980) 259–268.

D.M. Titterington and A.W. Bowman, A comparative study of smoothing procedures for ordered categorical data, *J. Statist. Comput. Simul.* **21** (1985) 291–312.