LOG-LINEAR MODELLING OF PAIRWISE INTEROBSERVER AGREEMENT ON A CATEGORICAL SCALE

MARK P. BECKER

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.

AND

ALAN AGRESTI

Department of Statistics, University of Florida, Gainesville, FL 32611, U.S.A.

SUMMARY

This article uses log-linear models to describe pairwise agreement among several raters who classify a sample on a subjective categorical scale. The models describe agreement structure simultaneously for second-order marginal tables of a multidimensional cross-classification of ratings. Practical difficulties arise in fitting the models, because models refer to pairwise marginal tables of a very large and sparse table. A standard analysis that treats the marginal tables as independent yields consistent estimates of model parameters, but not of the covariance matrix of the estimates. We estimate the covariance matrix using the jackknife. We apply the models to describe agreement between evaluations made by seven pathologists of carcinoma *in situ* of the uterine cervix, using a five-level ordinal scale. Previous analyses showed differences among the pathologists in their pairwise levels of agreement, but we observe near homogeneity in the dependence structure of their ratings.

1. INTRODUCTION

Suppose several raters separately classify each member of a sample, using a categorical measurement scale. Many categorical scales are quite subjective, and reliability assessment depends on evaluation of agreement among the raters. Kraemer[1] and Verducci *et al.*[2] recently discussed basic issues in evaluating agreement, and described a variety of ways of measuring it. Several authors[3–5] have described multi-rater agreement using generalizations of Cohen's[6] kappa. Others[7–9] showed difficulties with summarizing pairwise agreement by a single measure, and instead proposed *modelling* the structure of agreement among raters. This is also the approach taken in this article.

We illustrate agreement modelling using Table I, based on data presented in Landis and Koch[4] and originally reported by Holmquist *et al.*[10] Seven pathologists classified each of 118 slides in terms of carcinoma *in situ* of the uterine cervix, based on the most involved lesion, using the ordered categories: (1) negative; (2) atypical squamous hyperplasia; (3) carcinoma *in situ*; (4) squamous carcinoma with early stromal invasion; (5) invasive carcinoma. A 5^7 contingency table summarizes the joint classifications of the pathologists.

0277–6715/92/010101–14$07.00
© 1992 by John Wiley & Sons, Ltd.

Received June 1990
Revised April 1991

Table I. Cross-classification of seven pathologists on five categories

Pathologist ratings							Count	Pathologist ratings							Count	Pathologist ratings							Count
A	B	C	D	E	F	G		A	B	C	D	E	F	G		A	B	C	D	E	F	G	
1	1	1	1	1	1	1	10	2	3	2	2	2	1	2	1	3	3	3	4	3	2	3	1
1	1	1	1	2	1	1	8	2	3	2	2	3	1	3	1	3	3	3	4	3	2	4	1
1	1	2	1	1	1	1	2	2	3	2	2	3	2	2	1	4	2	3	2	3	2	3	1
1	1	2	1	2	1	1	2	2	3	2	2	3	2	3	2	4	3	1	1	2	1	2	1
1	2	1	1	1	1	1	1	2	3	2	2	4	1	2	1	4	3	1	3	3	2	3	1
1	2	2	1	2	1	2	1	2	3	2	2	4	1	3	1	4	3	3	2	3	2	3	1
1	3	2	1	2	1	1	1	3	2	2	2	2	1	1	1	4	3	3	3	3	2	3	2
1	3	2	2	2	1	2	1	3	2	2	2	2	1	2	1	4	3	3	3	3	3	3	3
2	1	1	1	2	1	1	1	3	3	2	1	3	2	2	1	4	3	3	3	3	5	3	1
2	1	1	2	1	1	1	1	3	3	2	2	2	1	2	1	4	3	3	4	3	3	3	2
2	1	2	1	1	1	1	1	3	3	2	2	2	2	3	1	4	3	3	4	4	3	3	1
2	1	2	1	2	1	1	1	3	3	2	2	3	1	3	4	4	3	4	2	3	3	3	1
2	1	2	2	2	1	2	1	3	3	2	2	3	2	2	1	4	3	4	2	4	1	3	1
2	2	1	1	2	1	1	1	3	3	2	2	3	2	3	2	4	4	3	2	4	1	3	2
2	2	1	1	2	1	2	1	3	3	2	2	3	3	3	1	4	4	3	3	4	3	3	1
2	2	1	2	2	1	2	1	3	3	2	2	4	2	3	1	4	4	3	4	4	3	4	1
2	2	2	1	1	1	2	1	3	3	2	3	2	2	3	1	4	4	4	2	4	3	3	1
2	2	2	1	2	2	2	1	3	3	2	3	3	1	3	1	4	4	4	2	5	1	3	1
2	2	2	2	3	1	2	2	3	3	2	3	3	3	3	2	4	4	4	3	3	3	3	1
2	3	1	1	2	1	1	1	3	3	3	2	3	1	3	2	5	3	3	2	3	2	3	1
2	3	1	1	2	1	2	1	3	3	3	2	3	2	3	3	5	3	3	3	4	1	3	1
2	3	1	1	3	1	1	1	3	3	3	2	3	3	3	1	5	3	4	2	3	4	3	1
2	3	1	2	3	1	3	1	3	3	3	2	4	2	3	2	5	5	1	4	5	5	4	1
2	3	2	1	3	2	2	1	3	3	3	2	4	3	3	1	5	5	5	4	5	5	5	1
2	3	2	2	2	1	3	1	3	3	3	3	3	2	3	5	5	5	5	5	5	5	5	1
2	3	2	2	2	2	2	1	3	3	3	3	3	3	3	4								

Data from Holmquist *et al.*,[10] as reported by Landis and Koch[4]

One possible analysis of these data compares the single-rater (that is, first-order) marginal distributions of the responses. For instance, we could analyse whether some pathologists tended to make higher ratings than other pathologists. Cox *et al.*[11] and Agresti *et al.*[12] conducted such an analysis and found strong evidence of differences. Though this is an informative comparison, it is not the focus of our attention. Instead we model the structure of agreement, focusing on second-order dependence in the joint distribution. An analysis of agreement should study this dependence as well as the marginal distributions, since there can be weak joint agreement in ratings even though the marginal distributions are similar; for instance, pairwise ratings could be statistically independent even though the first-order marginal distributions are identical. Using estimation methods discussed in this article, we can investigate whether all pairs of raters have the same structural pattern for agreement, and whether the raters have the same aggregate level of agreement with other raters.

Section 2 reviews log-linear models for two-rater agreement that this article generalizes. Section 3 analyses nominal-scale and ordinal-scale multi-rater agreement by simultaneously modelling marginal two-rater agreements. Multi-rater tables are often very large and sparse, and one must use special methods to fit the models and estimate the covariance matrix of parameter estimates. Section 4 uses models to analyse agreement structure in Table I. Section 5 gives ways of modelling agreement between several raters and a standard rating. The final section describes problems for future research that were suggested by our work.

2. MODELS FOR AGREEMENT

We first review log-linear models of agreement for the two-rater case. Suppose each rater separately classifies each of n subjects according to a fixed categorical scale. Let π_{ij} denote the probability of response i by the first rater and response j by the second rater. The general log-linear model for this case is

$$\pi_{ij} = \exp(\lambda_{i1} + \lambda_{j2} + \kappa_{ij})/[\Sigma_a \Sigma_b \exp(\lambda_{a1} + \lambda_{b2} + \kappa_{ab})]$$

where model parameters satisfy constraints such as $\Sigma\lambda_{i1} = \Sigma\lambda_{j2} = \Sigma_i \kappa_{ij} = \Sigma_j \kappa_{ij} = 0$. We will use the equivalent simpler formula

$$\log m_{ij} = \mu + \lambda_{i1} + \lambda_{j2} + \kappa_{ij} \qquad (1)$$

for the expected frequencies $\{m_{ij} = n\pi_{ij}\}$, where μ is a normalizing constant, satisfying $\exp(-\mu) = [\Sigma\Sigma \exp(\lambda_{a1} + \lambda_{b2} + \kappa_{ab})]/n$.

Darroch and McCloud[9] showed that under certain rather weak assumptions, agreement models satisfy the quasi-symmetry condition $\kappa_{ij} = \kappa_{ji}$ for all i and j. They also showed that the odds ratios

$$\tau_{ij} = (m_{ii} m_{jj})/(m_{ij} m_{ji}), \quad \text{for all } i < j,$$

are natural ones for interpreting models of agreement. Conditional on the event that the raters classify two subjects in categories i and j, τ_{ij} represents the odds that the ratings are concordant (that is, that the ratings agree for each subject) rather than discordant. The degree of agreement increases as $\{\tau_{ij}\}$ increase. Darroch and McCloud defined categories i and j as *indistinguishable* if $\tau_{ij} = 1$ and if $\tau_{ik} = \tau_{jk}$ for all other categories k. They used $\gamma_{ij} = 1 - (\tau_{ij})^{-1}$ to measure the *degree of distinguishability* of those categories.

To describe nominal-scale agreement, Tanner and Young[7] used quasi-independence models, such as log-linear model (1) with

$$\kappa_{ij} = \delta I(i = j) \qquad (2)$$

where $I(.)$ is the indicator function, equalling 1 when the raters agree and 0 otherwise. For this model, $\log(\tau_{ij}) = 2\delta$, and $\gamma_{ij} = 1 - \exp(-2\delta)$ describes beyond-chance agreement – that is, agreement beyond what we would expect if the first rating were statistically independent of the second.

For ordinal rating scales, given that two raters disagree, it is unrealistic to expect independent ratings. There is usually a moderate to strong positive association between the ratings, not confined to the main diagonal of the table. For instance, suppose there is an underlying continuum for the rating scale, for which the joint distribution of the raters' evaluations is bivariate normal. Then the discretized association will have approximately linear-by-linear form[13,14]

$$\kappa_{ij} = \beta u_i u_j \qquad (3)$$

for some set of monotone scores $\{u_i\}$. It follows from Lauritzen and Wermuth[15] that a latent structure model also implies this form of association when (i) the ratings are conditionally independent given a latent variable X, and (ii) X has a normal distribution with additive rating effects, conditional on the two ratings.

For model (3), $\log \tau_{ij} = (u_j - u_i)^2 \beta$. The degree of agreement increases as β increases, in the sense that the odds of concordant ratings for a pair of subjects increase in β. Since $\gamma_{ij} = 1 - \exp[-\beta(u_j - u_i)^2]$, the scores in model (3) help determine the distinguishability of categories. For a given location and scale constraints for the scores, the distinguishability of categories i and j increases as $|u_i - u_j|$ increases; if $u_i = u_j$, the conditional distribution for ratings by one rater is the same whether the other rater picks category i or j, and the categories are indistinguishable.

Agresti[16] described ordinal agreement patterns using model (3) and more general ones that permit extra agreement on the main diagonal. Becker[17] discussed related models. One can select the scores in these models *a priori*, or treat them as parameters and estimate them with sample data. For equal-interval scores, (3) is the uniform association model,[18] for which all local odds ratios $\{m_{ij}m_{i+1,j+1}/m_{i,j+1}m_{i+1,j}\}$ are identical.

3. MODELS FOR MARGINAL AGREEMENT AMONG SEVERAL RATERS

When d raters classify each subject on a scale with r categories, a contingency table having n observations in r^d cells describes the joint distribution of the sample ratings. It is sensible to construct a model that simultaneously describes agreement for the $\binom{d}{2}$ pairs of raters. Let $m_{(ab)ij}$ denote the expected frequency for the cell in row i and column j of the second-order marginal table for raters a and b. The log-linear model

$$\log m_{(ab)ij} = \mu_{(ab)} + \lambda_{(ab)1i} + \lambda_{(ab)2j} + \kappa_{(ab)}(i,j), \quad 1 \leqslant a < b \leqslant d \tag{4}$$

applies simultaneously to the $\binom{d}{2}$ second-order marginal distributions of the r^d contingency table. The model of independence for each pair of raters is the special case in which all κ parameters equal zero.

One would normally select the form of the κ parameter to reflect the nature of the response classification. When the rating scale is nominal, model (4) with the simple structure

$$\kappa_{(ab)}(i,j) = \delta_{(ab)}I(i=j) \tag{5}$$

usually fits much better than the model of independence; when the classification is ordinal, a dependence term of the form

$$\kappa_{(ab)}(i,j) = \beta_{(ab)}u_i u_j \tag{6}$$

is usually more appropriate.

More general forms for κ are also possible. For instance, one could let $\{\delta_{(ab)}\}$ in (5) vary by rating category. Often, though, one prefers to achieve parsimony by searching for similarity in the agreement structure among the marginal tables. Also, it is useful to summarize the agreement of each rater with the other raters, to check whether certain raters tend to have notably high or low aggregate levels of agreement. For models (5) and (6), one could attempt to describe aggregate agreement for each rater by considering the parsimonious special cases

$$\delta_{(ab)} = (\delta_a + \delta_b)/2 \tag{7a}$$

$$\beta_{(ab)} = (\beta_a + \beta_b)/2. \tag{7b}$$

Models (7*a*) and (7*b*) are equivalent to (5) and (6) when there are only $d = 2$ or 3 raters, but are simpler when $d > 3$. The rationale for such models is similar to that of two-way ANOVA models without interaction. They express the pairwise associations as a sum of main effects. There is also a resemblance to the Bradley–Terry model (see, for example, Agresti,[19] Section 10.6), in the sense that they describe parsimoniously $\binom{d}{2}$ parameters for pairs of raters by a single parameter for each rater.

When (7) holds, a large value of δ_a or β_a means that rater a tends to have strong agreements with the other raters. When it holds with $\delta_1 = \ldots = \delta_d$ or $\beta_1 = \ldots = \beta_d$, there is homogeneity in the agreement parameters for all $\binom{d}{2}$ second-order marginal tables. Alternatively, one could consider simpler versions of models (5) and (6) that have homogeneity within and between subsets of raters; for instance, $\delta_{(ij)}=\delta_1$ for i and j in subset C_1, $\delta_{(ij)}=\delta_2$ for i and j in subset C_2, and $\delta_{(ij)}=\delta_3$ for i in C_1 and j in C_2.

It is a non-trivial matter to fit these models, since they differ from ordinary log-linear models. They apply to second-order marginal distributions rather than to the interior cells of the complete r^d table to which the usual multinomial sampling model applies. One fitting approach is based on maximizing a multinomial likelihood for the complete table subject to the constraints that second-order marginal tables simultaneously satisfy (4). Haber[20] used the method of constrained maximum likelihood, described by Aitchison and Silvey,[21] to fit simple log-linear models to marginal configurations of contingency tables. The iterative procedure, however, involves inverting a matrix of rank greater than the number of cells in the table. With current computing capabilities, this approach is not feasible for large tables such as occur when there are several raters and the rating scale is polytomous.

We used an alternative, much simpler approach for fitting multi-rater agreement models. We fitted the models directly to the interior cells of a $r \times r \times \binom{d}{2}$ table, in which the kth $r \times r$ layer refers to classifications for the kth pair of raters; that is, each layer of this table is a second-order marginal table of the complete r^d table. This table sacrifices information about the joint ratings, retaining only the pairwise rating information. When d is large, this 'pairwise-ratings table' is much smaller and less sparse than the complete table. For Table I, for instance, the complete table has 118 observations in $5^7 = 78{,}125$ cells, whereas the pairwise-ratings table has $118 \times \binom{7}{2} = 2478$ observations in $5 \times 5 \times \binom{7}{2} = 525$ cells. We fitted the agreement model to the pairwise-ratings table by using a maximum likelihood routine that treats the $\binom{d}{2}$ layers as independent multinomial samples. For the general heterogeneous model (4), this corresponds to fitting an agreement model separately to each second-order margin of the complete table. For simpler models such as (7), the analysis pools information from different layers to estimate the agreement parameters. In either case, assuming the model holds for the two-factor margins of the complete table, the consistency of the sample proportions in each marginal table ensures the consistency of these 'pseudo ML' estimates.

Of course, samples in different layers of the pairwise-ratings table are not truly independent, since each layer classifies the same subjects. The pseudo ML estimated covariance matrix of the parameter estimates obtained by treating the $\binom{d}{2}$ layers as independent is not appropriate, and it is necessary to obtain a separate estimated covariance matrix that takes the dependence into account. White[22] specified the correct form for the covariance matrix for misspecified likelihoods, and Lipsitz *et al.*[23] showed that one can obtain an asymptotically equivalent estimate with the jackknife technique. In each step, one fits the model to the pairwise-ratings table obtained after deleting an observation from the complete table. The jackknife procedure also yields alternative estimates of the parameters in the model of interest.

With the jackknife approach to estimation, we assume that cell counts in the second-order marginal tables have multinomial covariance structure, but we do not need to make assumptions about the distribution of the d-way joint ratings in the complete r^d table. Another advantage of the jackknife is that we only need to focus on the parameters of interest in estimating the covariance matrix. For large problems such as modelling Table I, considerable simplification results from the ability to ignore the extremely large number of nuisance parameters (for example, the single-factor terms in model (1)). For other advantages of the jackknife when there may be model misspecification, see Shao.[24] When there is a large number of non-empty cells, the jackknife becomes more time consuming and it may be more practical to use the bootstrap to estimate the covariance matrix.

Suppose we want to compare the fit of two models to the $\binom{d}{2}$ marginal tables. To do this, we cannot use the usual likelihood-ratio statistics, since our analysis does not yield maximized likelihoods for the two models (and even if we could obtain such complete-table maximized likelihoods and fitted values, the table is often so sparse that likelihood-ratio statistics would have dubious utility). A simpler way to compare models is to construct Wald statistics, with use of estimated parameters for the more complex model. The Wald statistics are simply the statistics used in weighted least squares (WLS) methodology (Grizzle *et al.*[25]) for testing the fit of the simpler model when it is expressed in terms of parameters of the more complex model. Our use of this methodology parallels the functional asymptotic regression methodology outlined by Imrey *et al.*,[26] who used WLS methods in combination with estimates obtained from a ML fit of a model.

To illustrate, suppose we want to compare (6) with a homogeneous agreement model in which $\{\beta_{(ab)}\}$ are identical. Equivalently, we can test the fit of the model

$$\beta_{(ab)} = \beta, \quad \text{for all pairs} \quad (ab),$$

assuming model (6) holds. Let $t = \binom{d}{2}$, and let $\mathbf{h}$ be the $(t-1) \times 1$ vector $\mathbf{h} = (\hat{\beta}_{(12)} - \hat{\beta}_{(13)}, \hat{\beta}_{(13)} - \hat{\beta}_{(14)}, \ldots)'$. Let $\mathbf{A}$ be the $(t-1) \times t$ matrix with elements $a_{ii} = 1$, $a_{i,i+1} = -1$, $a_{ij} = 0$ otherwise. Let $\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})$ denote the jackknife estimated covariance matrix of $\{\hat{\boldsymbol{\beta}}_{(ab)}\}$, and let $\mathbf{S} = \mathbf{A}\,\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})\mathbf{A}'$. The statistic $\mathbf{h}'\,\mathbf{S}^{-1}\mathbf{h}$ tests homogeneity of the agreement parameters. Under the null hypothesis of homogeneous agreement, this statistic has an asymptotic chi-squared distribution with $\binom{d}{2} - 1$ degrees of freedom.

4. EXAMPLE

To fit multi-rater agreement models to Table I, we used the GLIM system (Numerical Algorithms Group[27]), release 3·77. We used SAS (PROC IML) for Wald tests. We first consider the independence model, the heterogeneous diagonal parameter model (5), and the heterogeneous uniform association model ((6) with equal-interval scores). The pseudo ML fits for these models for the different layers of the pairwise-ratings table are identical to those obtained by separate fitting of the corresponding bivariate models (for example, (2) and (3)) to the second-order marginal tables. To describe their goodness-of-fit, we report in Table II the components of the likelihood-ratio goodness-of-fit, statistic G^2 for the separate layers. These values are useful mainly for comparative purposes, since even the second-order marginal tables are quite sparse, with typically lots of zeros and small counts and a few large counts. (We do not report the Pearson statistic, because its behaviour is known to be highly erratic when tables contain both large and very small counts.[28]) Table II shows that the independence model fits poorly. Addition of the main-diagonal parameter (model (5)) makes a considerable improvement. The improvement is substantially greater yet for the uniform association model. Comparison of this model to model (5) illustrates how, for ordinal data, models that assume quasi-independence are generally inadequate.

Though we should interpret results cautiously, the uniform association model fits the 21 marginal tables quite well. Inspection of residuals for marginal tables having relatively large G^2 values indicates that in most cases this is due to an observation in which one rater's classification differs substantially from the others. In particular, the observation with ratings (5, 5, 1, 4, 5, 5, 4) by the 7 pathologists seemed highly influential in the fitting process for tables involving rater C. Table II also reports G^2 values for model (6) with this observation deleted. For marginal tables involving rater C, the G^2 value is considerably reduced compared to the full data set. (The reduction is even more dramatic for the Pearson statistics. The sum of its values equals 6068·9 for

Table II. Goodness-of-fit statistics for models fitted to pairwise marginal tables

Pathologist pair	Independence (d.f. = 16)	(5) (d.f. = 15)	Unif. Assoc. (d.f. = 15)		Additive model
A–B	131·2	30·9	16·2	15·7*	15·7*
A–C	139·3	88·7	44·5	30·1	30·6
A–D	117·3	74·7	24·3	25·5	25·7
A–E	113·3	62·7	13·4	13·1	13·1
A–F	97·3	81·8	24·8	24·6	24·6
A–G	133·4	52·9	5·5	7·9	8·3
B–C	94·1	43·6	23·0	7·9	8·9
B–D	97·1	53·2	11·4	11·4	11·8
B–E	136·2	53·2	4·7	4·3	9·7
B–F	85·6	62·4	21·6	21·1	21·1
B–G	141·3	40·3	7·5	8·3	8·4
C–D	105·9	55·1	41·5	33·6	34·1
C–E	104·1	70·3	28·7	13·9	14·2
C–F	88·4	58·8	35·5	19·2	19·7
C–G	123·4	40·8	26·2	13·3	13·7
D–E	101·2	83·2	32·6	33·0	34·0
D–F	85·2	51·2	11·3	12·2	13·3
D–G	149·2	68·9	2·6	2·3	8·4
E–F	84·8	75·8	38·0	37·3	38·4
E–G	128·6	59·6	4·5	6·3	6·4
F–G	90·9	52·9	7·2	9·3	9·7

* Observation (5, 5, 1, 4, 5, 5, 4) deleted

the 21 original tables, and 527·3 with this observation deleted.) For all subsequent analyses, we deleted this observation. We did this mainly to make more meaningful comparisons of fits of models. We do not encourage deletion of observations as a general strategy, and our later substantive conclusions about agreement remain unaltered if we do not delete this observation. In summary, model (6) with equally-spaced scores seems to describe well the agreement structure among these seven raters.

Table III contains pseudo ML and jackknife parameter estimates for model (6). It also reports their jackknife estimated standard errors, which apply to both sets of estimates. Using the pseudo ML estimate as the initial estimate, the iterative process for fitting the model with a deleted observation converged in almost all instances within two steps. In each case, the jackknife parameter estimate is slightly weaker than the pseudo ML estimate, which suggests that the pseudo ML estimates may be biased upwards. We do not report the 21 × 21 jackknife correlation matrix of the estimates; the estimated correlation between estimates of $\beta_{(ab)}$ and $\beta_{(cd)}$ was almost always weak (below 0·20) when raters (a, b) and (c, d) formed disjoint sets, but ranged from weak to quite strong (several values exceeding 0·75) when the sets had a rater in common. For this heterogeneous agreement model, one could also use ML with separate marginal tables to obtain standard errors of the parameter estimates, but one would need the jackknife or some other method to estimate correlations of estimates from separate tables.

Inspection of Table III reveals a certain consistency of results for the 21 marginal tables. For each table, $\hat{\beta}$ is positive and indicates a relatively strong local odds ratio. It therefore makes sense to consider possible simplification of the agreement model. One can fit models that are special cases of (6) in two ways. As with (6), one can fit the model directly to the pairwise-ratings table to obtain pseudo ML estimates, and use the jackknife to obtain estimated standard errors, or, one

Table III. Parameter estimates for model (6) fitted to marginal tables (jackknife a.s.e. values in parentheses)

Pathologist	B	C	D	E	F	G
A	1·84* 1·73† (0·340)	1·88 1·75 (0·461)	1·49 1·42 (0·291)	1·53 1·47 (0·263)	1·15 1·08 (0·270)	2·18 2·04 (0·422)
B		1·67 1·61 (0.271)	1·68 1·62 (0·256)	2·73 2·59 (0·444)	1·34 1·24 (0·352)	2·56 2·36 (0·563)
C			1·53 1·45 (0·301)	1·81 1·75 (0·284)	1·42 1·34 (0·352)	2·29 2·16 (0·448)
D				1·32 1·27 (0·247)	1·41 1·33 (0·286)	3·88 3·37 (0·947)
E					0·91 0·84 (0·276)	2·48 2·30 (0·472)
F						1·79 1·64 (0·323)

* ML estimates
† Jackknife estimates

can apply weighted least squares, using as responses the pseudo ML estimates of $\{\beta_{(ab)}\}$ for model (6) and using the jackknife estimated covariance matrix of those estimates. The former approach gives fitted values in the cells of the marginal tables, but the second approach may give more efficient estimates of model parameters, since it recognizes the dependences among the marginal tables in forming the estimates of model parameters.

When we used the pseudo ML approach to fit the homogeneous agreement version of (6), which assumes all $\beta_{(ab)}$ equal some value β, we obtained $\bar{\beta} = 1{\cdot}70$ for the estimated common association. The asymptotic standard error, estimated with the jackknife, was a.s.e. $(\bar{\beta}) = 0{\cdot}15$. We compared the fit of the homogeneous uniform association model to the independence model using

$$z = \bar{\beta}/[\text{a.s.e.}(\bar{\beta})\} = 11{\cdot}28.$$

There is extremely strong evidence that the agreement is better than that expected by chance. The WLS estimate of β for the model $\beta_{(ab)} = \beta$ equals 1·60, with a.s.e = 0·12, leading to the same substantive conclusion.

Using the jackknife estimated covariance matrix of the β estimates from the heterogeneous uniform association model, we tested the adequacy of the simpler homogeneous model by conducting a Wald test of equality of the 21 β parameters from the marginal tables. For the pseudo ML parameter estimates, this statistic equalled 51·4, based on d.f. = 20. The heterogeneous model provided some improvement in fit. The estimated values in Table III, however, indicate that, except for perhaps the D–G and E–F agreements, none of the estimates differs substantively from the 'uniform' estimate.

A less drastic special case of model (6) is the additive model (7*b*). The pseudo ML approach gives fitted values in the cells of the pairwise-ratings table, and Table II shows components of the

Table IV. Pseudo ML estimated parameters $\{\hat{\beta}_a\}$ for additive model, and estimated covariances (× 1000)

				Rater			
	A	B	C	D	E	F	G
				Estimate			
Covariance with:	1·56	2·09	1·81	1·60	1·51	0·72	3·28
A	21·7	− 0·5	1·1	0·1	2·1	1·5	− 3·1
B	− 0·5	15·9	− 3·0	− 0·3	4·7	1·3	4·4
C	1·1	− 3·0	22·5	− 1·4	1·4	1·1	5·6
D	0·1	− 0·3	− 1·4	20·4	− 2·1	2·4	5·6
E	2·1	4·7	1·4	− 2·1	17·0	− 3·9	2·6
F	1·5	1·3	1·1	2·4	− 3·9	30·2	− 0·8
G	− 3·1	4·4	5·6	5·6	2·6	− 0·8	33·7

Table V. Fit provided by additive model (7*b*) to pseudo ML estimates of $\{\beta_{(ab)}\}$ in heterogeneous model (6)

Pathologist	B	C	D	E	F	G
A	1·84*	1·88	1·49	1·53	1·15	2·18
	1·83†	1·69	1·58	1·64	1·14	2·42
B		1·67	1·68	2·73	1·34	2·56
		1·96	1·84	1·80	1·40	2·68
C			1·53	1·81	1·42	2·29
			1·71	1·55	1·27	2·55
D				1·32	1·41	3·88
				1·55	1·15	2·43
E					0·91	2·48
					1·11	2·39
F						1·79
						2·00

* Model (6)
† Model (7*b*)

G^2 statistic for each pair of raters. This simple model seems to fit as well as the heterogeneous model (6), except for the D–G and B–E marginal tables. Table IV shows pseudo ML estimates of $\{\beta_a\}$, together with estimated covariances of those estimates. Table V compares the pseudo ML estimates of $\{\beta_{(ab)}\}$ for model (6) to those predicted by the additive model (7*b*). The additive model does a reasonably good job in predicting the heterogeneous estimates for 19 of the 21 pairs of raters, with marked discrepancies occurring for the D–G and B–E pairs. The WLS test of fit of the additive model (assuming (6) holds) has chi-squared statistic equal to 26·2, based on d.f. = 14, and reflects this lack of fit. Simpler models that assume homogeneity of pairwise agreement parameters within and between subsets of raters showed even greater lack of fit.

We can use the estimated parameters for the additive model to describe aggregate agreement for the raters. A Bonferroni multiple comparison of the 21 pairs of estimates of $\{\beta_a\}$, using an overall 0·95 confidence coefficient, reveals that the agreement component for rater G is significantly higher than all others, and the agreement component for rater F is significantly lower than all others. The only other slight evidence of a difference is between E and B. Thus, rater G tends to have highest agreement with other raters, rater F the lowest, and the similarity of other estimated $\{\beta_a\}$ suggests that agreements between other pairs of raters are approximately uniform.

In summary, we can give quite a simple description to Table I. The simple additive model (7*b*) fits reasonably well, rater G has highest and rater F has lowest aggregate agreement with the other raters, and the other raters have similar aggregate agreements and similar pairwise levels of agreement.

Schouten[5] analysed Table I by calculating Cohen's kappa for each pair of pathologists, using the collapsed scale in which he combined the first two and last three categories. Schouten concluded that agreement among pathologists A, B, E and G is substantial, each of these pathologists has at least moderate agreement with pathologist C, and pathologists D and F each has less agreement with the others. Schouten's conclusions reflect the variation in marginal distributions for the seven pathologists. Pathologists who had similar marginal distributions for this collapsed scale tended to have higher levels of agreement, as described by kappa.

Our conclusion differs somewhat from Schouten's, because our models use the full five-point ordinal scale and describe the agreement structure with *adjustment* for marginal discrepancies. The models describe agreement in terms of concordance of ratings, and there can be high concordance (for example, large values of τ_{ij}) even when there are substantial discrepancies in marginal distributions. In summary, though there are substantive differences in the seven marginal distributions of ratings, our analyses suggests that (adjusting for such differences) there is near homogeneity in the pattern and strength of agreement.

5. COMPARING SEVERAL RATERS TO A STANDARD

Simplifications in the fitting process are sometimes possible when we want to describe simultaneously agreement between d raters and a 'standard' rating. The standard might be a known correct classification, or it might simply be the current 'best' known way of making a rating. For nominal rating scales, Tanner and Young[7] considered two types of models for this situation – one when raters examine independent samples, and the other when raters examine the same sample. We now propose models for ordinal-scale agreement for these two cases, and we propose models of nominal-scale agreement for the second case that differ from the Tanner and Young model.

When raters examine independent samples, the responses form d separate tables, where the kth table compares rater k to the standard. Let m_{ijk} denote the expected frequency when the standard classification is i and the rater classification is j, for the kth rater, and consider the model

$$\log m_{ijk} = \mu + \lambda_i^{\mathrm{S}} + \lambda_j^{\mathrm{N}} + \lambda_k^{\mathrm{R}} + \lambda_{ik}^{\mathrm{SR}} + \lambda_{jk}^{\mathrm{NR}} + \kappa_k(i,j) \tag{8}$$

where S is standard rating, N is non-standard rating and R is rater.

For instance, the structure $\kappa_k(i,j) = \beta_k u_i u_j + \delta_k I(i=j)$ permits levels of association $\{\beta_k\}$ and main-diagonal elevation parameters $\{\delta_k\}$ between each rater and the standard to vary by rater. Special cases include heterogeneous linear-by-linear association for the S–N agreements ($\delta_1 = \ldots \delta_d = 0$) and the Tanner and Young model for nominal-scale agreement between each rater and the standard ($\beta_1 = \ldots = \beta_d = 0$). Since the k layers of this table now truly are independent samples, we can fit the model with standard methods for log-linear models.

Next, suppose the d raters examine the same sample. Let $\mathbf{i} = (i_0, \ldots, i_d)$, where i_0 is the rating by the standard. Let $\{m_{\mathbf{i}}\}$, with $\mathbf{i} = (i_0, \ldots, i_d)$, denote the expected frequencies in the r^{d+1} cross-classification of ratings. Let $m_{ij(k)}$ be the expected frequency in the cell in row i and column j for the marginal table between rater k and the standard, $k = 1, \ldots, d$. We could model the d

pairwise marginal agreements between the raters and the standard by

$$\log m_{ij(k)} = \mu_{(k)} + \lambda_{i(k)} + \lambda_{j(k)} + \kappa_k(i,j), \quad 1 \leqslant k \leqslant d, \tag{9}$$

where the form for $\kappa_k(i,j)$ depends on the measurement scale. We can fit such a model using the methodology described in Section 3, treating the data as d independent $r \times r$ tables for purposes of obtaining the estimates.

For the second case, it might sometimes be reasonable to assume that, conditional on the standard rating, ratings by other pairs of raters are statistically independent. Then model (9) is equivalent to the log-linear model

$$\log m_{\mathbf{i}} = \mu + \sum_{k=0}^{d} \lambda_{ki_k} + \sum_{k=1}^{d} \kappa_k(i_k, i_0) \tag{10}$$

for the complete table, since standard collapsibility conditions imply that the d partial associations of each rater with the standard are identical to the marginal associations. Thus, when the complete table is not too large, one can obtain parameter estimates and the estimated covariance structure for model (9) directly by fitting model (10). One can then check the conditional independence assumption by analysing the fit of (10). When the model fits adequately, one can investigate further whether a simpler model of homogeneous agreement with the standard fits adequately.

When all raters observe the same sample, Tanner and Young proposed a log-linear model that allows the most general interaction pattern among the raters and describes conditional agreement between each rater and the standard. A practical difficulty with this model is that it forces expected frequency estimates to equal the observed data in each cell of the r^d marginal cross-classification of the raters. Because of the sparseness of this table, ML estimates rarely exist for this appraoch.

6. COMMENTS

In general, a model that specifies homogeneous pairwise agreement structure may fit well even though raters have different marginal distributions and even though pairs of raters have different levels of agreement as measured by an index such as kappa. When this happens, it often indicates that variation in the index of agreement is due to the variation in the marginal distributions. If the observers could calibrate their ratings so that the marginal distributions were identical (perhaps matching some standard distribution), they might be interchangeable with respect to their distributions of pairwise ratings. For instance, for an ordinal rating scale, this interchangeability might occur if some observers adjusted their ratings upwards or downwards.

Agreement involves both similarity of first-order marginal distributions for raters and strong pairwise association between them. Neither strong association nor identical marginal distributions is sufficient to ensure strong agreement. We have focused on modelling association in this article. Strong association is indicative of consistency between classification of different raters, at least in terms of odds of concordance, and is the primary determinant of agreement with ratings calibrated to match a standard.

An alternative approach to modelling agreement is to focus on *conditional* agreement among the raters.[7] Model parameters then describe agreement between a pair of raters, *for fixed ratings by the other $d-2$ raters*. In most applications we believe this has less descriptive important than *marginal* agreement for each pair of raters. It is usually not sensible to condition on other ratings to describe the agreement between two raters.

The methods we have employed in this paper apply more generally than to rater agreement problems. In many longitudinal studies, marginal associations are more relevant than partial associations. In a panel study in which one observes responses for each subject at times $1, 2, \ldots, d$, one might want to describe simultaneously the associations between responses at all pairs of times. One might have interest in whether the association is homogeneous for all pairs of times that are the same number of time units apart, for instance, or, in a social mobility study, one might want to model simultaneously associations that correspond to one-step transitions.

For some data sets, alternative analyses to those we have described might be more appropriate. For instance, suppose there were two different methods of making ratings, and one sample of raters used one method and a second sample used the other method. One might have interest in whether there was a difference between the level of agreement among raters who used one method and the level of agreement among raters who used the other method. For model (6), one could describe aggregate agreement for the two methods using the special case of a homogeneous β for pairs of raters using method one, and another homogeneous β for pairs of raters using method two. One would then compare levels of agreement using the difference in β estimates. For this problem, a 'random effects' approach might be more reasonable. One could treat raters using each method as a sample, and use average pairwise agreement for each method to estimate a mean pairwise agreement for a population of raters.

More generally, one could incorporate covariates in the model. Interest can then focus additionally on association between each rater's rating and the covariates as well as on the agreement among raters, controlling for the covariates. One can view the analysis described in the previous paragraph as resulting from use of a single binary covariate that identifies the method of making the rating.

7. FUTURE RESEARCH PROBLEMS

Our work in this area suggests several interesting problems for future research. Many of these concern special difficulties presented by sparse data. For instance, in Section 4 we analysed goodness-of-fit by checking each two-way marginal table separately. It would be useful to test goodness-of-fit simultaneously for the 21 marginal tables. Using an estimate of the joint covariance matrix of the 525 cell counts in those marginal tables, one could obtain an approximate distribution of X^2 or G^2 as calculated for the pairwise-rating table. It follows from results in Rao and Scott[29] that the asymptotic distribution is a weighted sum of 315 χ_1^2 random variables, where the weights are eigenvalues of a 315×315 matrix. The application of these results to highly sparse tables such as in this paper is dubious, and this is a topic for future work. The same remark applies to use of statistics such as G^2 to compare fits of nested, unsaturated models for the pairwise-rating table. Rotnitzky and Jewell[30] have conducted some work of this type, in a simpler setting, for semiparametric generalized linear models for cluster correlated data.

Another concern relates to the efficiency of treating the second-order marginal tables as independent in estimating parameters in agreement models. We used this procedure because of its simplicity compared to maximizing the likelihood for the entire table. Based on efficiency results in Liang and Zeger[31] for cases in which each subject has the same number of repeated observations, we conjecture that there is little efficiency loss when the jackknife estimates from separate marginal tables have weak to moderate correlations, as was the case in the example we analyzed.

ACKNOWLEDGEMENTS

Dr. Becker's research was partially supported by grants CA-53787 and CA-09168 from the National Cancer Institute and grant SES-8618207 from the National Science Foundation, and

Dr. Agresti's research was partially supported by grant GM 43824 from the National Institutes of Health. The authors thank the editor and the referees for their helpful comments.

REFERENCES

1. Kraemer, H. C. 'Assessment of 2 × 2 associations: generalization of signal-detection methodology', *American Statistician*, **42**, 37–49 (1988).
2. Verducci, J. S., Mack, M. E. and DeGroot, M. H. 'Estimating multiple rater agreement for a rare diagnosis', *Journal of Multivariate Analysis*, **27**, 512–535 (1988).
3. Fleiss, J. L. 'Measuring nominal scale agreement among many raters', *Psychological Bulletin*, **76**, 378–382 (1971).
4. Landis, J. R. and Koch, G. G. 'An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers', *Biometrics*, **33**, 363–374 (1977).
5. Schouten, H. J. A. 'Measuring pairwise interobserver agreement when all subjects are judged by the same observers', *Statistica Neerlandica*, **36**, 45–61 (1982).
6. Cohen, J. 'A coefficient of agreement for nominal scales', *Educational and Psychological Measurement*', **20**, 37–46 (1960).
7. Tanner, M. A. and Young, M. A. 'Modeling agreement among raters', *Journal of the American Statistical Association*, **80**, 175–180 (1985).
8. Jørgensen, B. 'Estimation of interobserver variation for ordinal rating scales', *in Proceedings of GLIM Conference*, *Lancaster*, Springer-Verlag, Berlin, 1985.
9. Darroch, J. N. and McCloud, P. I. 'Category distinguishability and observer agreement', *Australian Journal of Statistics*, **28**, 371–388 (1986).
10. Holmquist, N. S., McMahon, C. A. and Williams, O. D. 'Variability in classification of carcinoma in situ of the uterine cervix', *Archives of Pathology*, **84**, 334–345 (1967).
11. Cox, M. A. A., Przepiora, P. and Plackett, R. L. 'Multivariate contingency tables with ordinal data', *Utilitas Mathematica*, **21A**, 29–42 (1982).
12. Agresti, A., Lipsitz, S. and Lang, J. 'Comparing marginal distributions of large, sparse contingency tables', *Computational Statistics and Data Analysis*, (1992) to appear.
13. Goodman, L. A. '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', *Annals of Statistics*, **13**, 10–69 (1985).
14. Becker, M. 'On the bivariate normal distribution and association models for ordinal categorical data', *Statistics & Probability Letters*, **8**, 435–440 (1989).
15. Lauritzen, S. L. and Wermuth, N. 'Graphical models for association between varibles, some of which are qualitative and some quantitative', *Annals of Statistics*, **17**, 31–57 (1989).
16. Agresti, A. 'A model for agreement between ratings on an ordinal scale', *Biometrics*, **44**, 539–548 (1988).
17. Becker, M. 'Association models to analyse agreement data: two examples', *Statistics in Medicine*, **8**, 1199–1208 (1989).
18. Goodman, L. A. 'Simple models for the analysis of association in cross-classifications having ordered categories', *Journal of the American Statistical Association*, **74**, 537–552 (1979).
19. Agresti, A. *Categorical Data Analysis*, Wiley, New York, 1990.
20. Haber, M. 'Log-linear models for correlated marginal totals of a frequency table', *Communications in Statistics, Theory and Methods*, **14**, 2845–2856 (1985).
21. Aitchison, J. and Silvey, S. D. 'Maximum-likelihood estimation of parameters subject to restraints', *Annals of Mathematical Statistics*, **29**, 813–828 (1958).
22. White, H. 'Maximum likelihood estimation of misspecified models', *Econometrica*, **50**, 1–25 (1982).
23. Lipsitz, S. R., Laird, N. M. and Harrington, D. P. 'Using the jackknife to estimate the variance of regression estimators from repeated measures studies', *Communications in Statistics*, **19**, 821–845 (1990).
24. Shao, J. 'Jackknifing in generalized linear models', Technical Report STAT-89-21, Department of Statistics and Actuarial Science, University of Waterloo, 1989.
25. Grizzle, J. E., Starmer, C. F. and Koch, G. G. 'Analysis of categorical data by linear models', *Biometrics*, **25**, 489–504 (1969).
26. Imrey, P. B., Koch, G. G. and Stokes, M. E. 'Categorical data analysis: some reflections on the log linear model and logistic regression. Part I: Historical and methodological overview', *International Statistical Review*, **49**, 265–283 (1981).

27. Numerical Algorithms Group. *The GLIM Release 3.77 Manual*, Numerical Algorithms Groups Inc., Downers Grove, Illinois, 1985.
28. Haberman, S. J. 'A warning on the use of chi-squared statistics with frequency tables with small expected cell counts', *Journal of the American Statistical Association*, **83**, 555–560 (1988).
29. Rao, J. N. K. and Scott, A. J. 'On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data', *Annals of Statistics*, **12**, 46–60 (1984).
30. Rotnitzky, A. and Jewell, N. P. 'Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data', *Biometrika*, **77**, 485–497 (1990).
31. Liang, K.-Y. and Zeger, S. L. 'Longitudinal data analysis using generalized linear models', *Biometrika*, **73**, 13–22 (1986).