Quasi-Symmetric Latent Class Models, with Application to Rater Agreement
Author(s): Alan Agresti and Joseph B. Lang
Reviewed work(s):
Source: *Biometrics*, Vol. 49, No. 1 (Mar., 1993), pp. 131-139
Published by: International Biometric Society
Stable URL: http://www.jstor.org/stable/2532608
Accessed: 30/04/2012 16:21

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

International Biometric Society is collaborating with JSTOR to digitize, preserve and extend access to *Biometrics.*

http://www.jstor.org

BIOMETRICS 49, 131–139
March 1993

Quasi-Symmetric Latent Class Models, with Application to Rater Agreement

Alan Agresti and Joseph B. Lang*

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

SUMMARY

Suppose we observe responses on several categorical variables having the same scale. We consider latent class models for the joint classification that satisfy quasi-symmetry. The models apply when subject-specific response distributions are such that (i) for a given subject, responses on different variables are independent, and (ii) odds ratios comparing marginal distributions of the variables are identical for each subject. These assumptions are often reasonable in modeling multirater agreement, when a sample of subjects is rated independently by different observers. In this application, the model parameters describe two components of agreement—strength of association between classifications by pairs of observers and degree of heterogeneity among the observers' marginal distributions. We illustrate the models by analyzing a data set in which seven pathologists classified 118 subjects in terms of presence or absence of carcinoma, yielding seven categorical classifications with the same binary scale. A good-fitting model has a latent classification that differentiates between subjects on whom there is agreement and subjects on whom there is disagreement.

1. Introduction

Latent class models express the joint distribution among a set of categorical variables as a mixture of distributions, each component of which satisfies mutual independence among the variables. Each distribution in the mixture applies to a cluster of subjects representing a separate class of a categorical latent variable, those subjects being homogeneous in some sense. Since Goodman's (1974) development of maximum likelihood (ML) procedures for fitting latent class models, they have been used for a wide variety of applications. For instance, Clogg (1981) used them to analyze intergenerational mobility, interpreting the latent classes as different social classes. Aitkin, Anderson, and Hinde (1981) used them to analyze educational research data by clustering teachers into distinct teaching styles. Latent class models have also been used to assess agreement and disagreement among subjects' responses to several survey items (Clogg, 1979) or among ratings by several judges (Aickin, 1990; Dillon and Mulani, 1984; Espeland and Handelman, 1989; Uebersax and Grove, 1990). See Goodman (1974), Haberman (1979, Chap. 10), and McCutcheon (1987) for introductions to latent class models.

This article discusses a latent class model in which each observed variable has the same categorical scale, and the relationship among those variables satisfies quasi-symmetry. Such models are appropriate when subject-specific response distributions satisfy two basic assumptions. The first is a local independence assumption, whereby for a given subject, responses on different variables are independent. The second is a lack of interaction assumption, whereby odds ratios comparing marginal distributions of observed variables are identical for each subject. Latent classes in the models consist of sets of subjects who are homogeneous in terms of having the same response distributions. The proposed quasi-symmetric latent class models are parsimonious, having identical associations between each observed variable and the latent variable. For ordinal variables, even simpler models are relevant, such as one having a common linear-by-linear association between each observed variable and the latent variable.

As mentioned above, several authors have used latent class models to investigate interrater agreement. Other authors, including Agresti (1988), Becker (1990), and Darroch and McCloud (1986), have used quasi-symmetric models for this purpose. In this article we combine the approaches and use quasi-symmetric latent class models to analyze agreement. To illustrate the models, we analyze Table 1, based on data presented by Landis and Koch (1977). Seven pathologists classified each of

* *Current address*: Department of Statistics, University of Iowa, Iowa City, Iowa 52242.

Key words: Association models; EM algorithm; Loglinear models; Marginal homogeneity; Ordinal variables; Rasch model.

Table 1
Diagnoses of carcinoma (1 = *no*, 2 = *yes*)

Pathologist							
A	*B*	*C*	*D*	*E*	*F*	*G*	Count
1	1	1	1	1	1	1	34
1	1	1	1	2	1	1	2
1	2	1	1	1	1	1	6
1	2	1	1	1	1	2	1
1	2	1	1	2	1	1	4
1	2	1	1	2	1	2	5
2	1	1	1	1	1	1	2
2	1	2	1	2	1	2	1
2	2	1	1	1	1	1	2
2	2	1	1	1	1	2	1
2	2	1	1	2	1	1	2
2	2	1	1	2	1	2	7
2	2	1	1	2	2	2	1
2	2	1	2	1	1	2	1
2	2	1	2	2	1	2	2
2	2	1	2	2	2	2	3
2	2	2	1	2	1	2	13
2	2	2	1	2	2	2	5
2	2	2	2	2	1	2	10
2	2	2	2	2	2	2	16
							118

118 slides in terms of carcinoma *in situ* of the uterine cervix. Category 2 represents a diagnosis of carcinoma. The data have been analyzed using kappa-type measures of agreement by those authors and by Schouten (1982), and using loglinear models by Becker and Agresti (1992). The data consist of 118 observations in a cross-classification of the ratings having $2^7 = 128$ cells. A quasi-symmetric model with three latent classes provides simple interpretations for pairwise agreement structure among the seven raters. We use model parameters to describe strength of association between ratings as well as degree of heterogeneity among the raters' marginal distributions on the binary scale. The three latent classes correspond to subjects for whom raters generally agree on the presence of carcinoma, subjects for whom raters generally agree on the absence of carcinoma, and subjects for whom there is strong disagreement.

Section 2 describes the basic assumptions and introduces the concepts of quasi-symmetry and local independence. Section 3 defines quasi-symmetric latent class models, and Section 4 presents a simpler model for ordinal scales. Section 5 discusses model fitting and inference, and Section 6 applies models to Table 1. The final section comments on the scope of the models and relates them to Rasch models and Rasch mixture models proposed by Lindsay, Clogg, and Grego (1991).

2. Quasi-Symmetry and Local Independence

Suppose we observe responses on R categorical variables that have the same set of I categories. We observe the variables for n subjects, randomly sampled from a population of S subjects. We permit subject-specific variability in response distributions. For subject s and variable r, let ϕ_{sri} denote the probability of response in category i. To simplify notation in the following discussion, we illustrate models for $R = 3$, with variables denoted by A, B, and C. The models extend in an obvious manner to arbitrary integer $R \geqslant 2$.

For a given subject s, we assume that classifications on different variables are statistically independent. That is, the probability that subject s has responses h on variable A, i on variable B, and j on variable C, equals $\phi_{s1h}\phi_{s2i}\phi_{s3j}$. Letting π_{hij} denote the probability of these three outcomes for a randomly selected subject, we have

$$\pi_{hij} = S^{-1} \sum_{s} \phi_{s1h}\phi_{s2i}\phi_{s3j}. \tag{2.1}$$

Associations among variables in the $\{\pi_{hij}\}$ distribution are due to heterogeneity among subjects in their response distributions. The data to be analyzed consist of sample cell counts $\{n_{hij}\}$ specifying frequencies for the I^3 possible combinations of outcomes.

We also assume that $\{\phi_{sri}\}$ satisfy the condition of no three-factor interaction; that is, the association between item observed and response is the same for each subject, so ϕ_{sri} has the form

$$\phi_{sri} = \alpha_{sr}\beta_{si}\gamma_{ri}. \tag{2.2}$$

As we discuss in Section 7, this means that $\{\phi_{sri}\}$ satisfy a generalized Rasch model. Equivalently, (2.2) means that the signals emitted by the subjects rated and the rater differences combine without interaction in affecting the response. Darroch and McCloud (1986) gave arguments supporting assumption (2.2) for modeling rater agreement on subjective categorical scales. They also noted by substituting (2.2) into (2.1) that $\{\pi_{hij}\}$ consequently satisfy quasi-symmetry (Caussinus, 1965); that is, π_{hij} has the form

$$\pi_{hij} = a_h b_i c_j d_{hij}, \tag{2.3}$$

where d_{hij} is identical for every permutation of the subscripts. Darroch and McCloud (1986) argued that models for agreement should satisfy quasi-symmetry.

Next, suppose there is a categorical variable X, having L levels, such that subjects in each level of X are homogeneous; namely, for each l, r, and i ($l = 1, \ldots, L$; $r = 1, \ldots, R$; $i = 1, \ldots, I$), suppose there is a probability ρ_{lri} such that for all subjects s in category l of X, $\phi_{sri} = \rho_{lri}$. When $\{\phi_{sri}\}$ satisfy no three-factor interaction, then so do $\{\rho_{lri}\}$. Variable X is unobserved, hence a latent variable. The assumption of independence of the observed variables within each level of X is referred to as "local independence."

For a randomly selected subject, let π_{hijl} denote the probability of outcomes (h, i, j) for variables (A, B, C), and categorization in class l of X. Then π_{hij} in (2.1) satisfies $\pi_{hij} = \pi_{hij+}$, where the subscript $+$ denotes summation over that index, and $\rho_{l1h} = \pi_{h++l}/\pi_{+++l}$, $\rho_{l2i} = \pi_{+i+l}/\pi_{+++l}$, $\rho_{l3j} = \pi_{++jl}/\pi_{+++l}$. Since

$$\pi_{hijl} = \pi_{+++l}\rho_{l1h}\rho_{l2i}\rho_{l3j},$$

$\{\pi_{hijl}\}$ satisfy mutual independence of A, B, and C, given X; that is, they satisfy the loglinear model for which the sufficient marginal configurations are represented by the notation (AX, BX, CX). The loglinear model representation (Haberman, 1979) is equivalent to the standard probabilistic latent class model for three observed variables and a single latent variable.

For a random sample of n subjects, let $\{m_{hijl} = n\pi_{hijl}\}$ denote expected frequencies for the unobserved A–B–C–X cross-classification. Formula (2.3) suggests that it may be fruitful to consider models for which $\{m_{hij+}\}$ satisfy quasi-symmetry.

3. A Simplified Latent Class Model

The loglinear version (AX, BX, CX) of the ordinary latent class model for $\{m_{hijl}\}$ corresponds to the nonlinear model for the expected frequencies $\{m_{hij} = m_{hij+}\}$ of observed cells having form

$$\log m_{hij} = \mu + \lambda_h^A + \lambda_i^B + \lambda_j^C + \log\left[\sum_l \exp(\lambda_l^X + \lambda_{hl}^{AX} + \lambda_{il}^{BX} + \lambda_{jl}^{CX})\right]. \tag{3.1}$$

This model satisfies quasi-symmetry if the term in brackets in (3.1) is the same for every permutation of (h, i, j). But this condition is equivalent to

$$\lambda_{il}^{AX} = \lambda_{il}^{BX} = \lambda_{il}^{CX} \quad \text{for all } i \text{ and } l. \tag{3.2}$$

So, no three-factor interaction for $\{\rho_{lri}\}$ implies highly parsimonious models having identical association between each observed variable and the latent variable. We refer to model (3.1) with condition (3.2) as the *quasi-symmetric latent class* (QLC) *model.*

We first consider some implications of the QLC model when $L = I = 2$. Without loss of generality, we scale the parameters describing the association of each observed variable with X so that $\lambda_{12} = \lambda_{21} = \lambda_{22} = 0$. Let λ denote the common value of $\{\lambda_{11}\}$, and let $I(\cdot)$ denote the indicator function. When the QLC model holds, the logit model

$$\text{logit}[\Pr(X = 1 \mid A = h, B = i, C = j)] = \alpha + \lambda[I(h = 1) + I(i = 1) + I(j = 1)] \tag{3.3}$$

describes effects of the observed variables on the latent variable. That is, the odds of response 1 equal $\exp(\alpha)$ when $A = B = C = 2$, and they are multiplied by $\exp(t\lambda)$ when t observed variables equal 1. Standard conditions are satisfied for collapsing over levels of observed variables, implying that the marginal relationship between each observed variable and X has odds ratio $\exp(\lambda)$.

For the QLC model, one can make simple comparisons of marginal distributions of the observed variables within each latent class, using odds ratios of $\{\rho_{lri}\}$. For instance, for variables A and B when

$I = 2$, we would use the odds ratios

$$\frac{\Pr(A = 1 \mid X = l)/\Pr(A = 2 \mid X = l)}{\Pr(B = 1 \mid X = l)/\Pr(B = 2 \mid X = l)} = \frac{\rho_{l11}/\rho_{l12}}{\rho_{l21}/\rho_{l22}} = \exp(\delta_A - \delta_B), \quad l = 1, \ldots, L, \tag{3.4}$$

where $\delta_A = \lambda_1^A - \lambda_2^A, \ldots, \delta_C = \lambda_1^C - \lambda_2^C$. Since $\{\rho_{lri}\}$ satisfy no three-factor interaction, the odds ratios are identical for every latent class l.

In every level of X, each subject has an observation for each observed variable. The independence of the subject-by-item table implies that (collapsing over subjects) the marginal probabilities $\{\pi_{1++}\}$, $\{\pi_{+1+}\}$, and $\{\pi_{++1}\}$ have the same ordering as $\{\delta_A, \delta_B, \delta_C\}$. When $\delta_A = \delta_B = \delta_C$ in the QLC model, in each latent class the observed variables have identical response distributions. The A–B–C contingency table then satisfies first-order marginal homogeneity, and the QLC model satisfies complete symmetry. In practice, we rarely expect this special case of the model to fit well.

4. An Ordinal Quasi-Symmetric Latent Class Model

When the observed categorical scale is ordinal, one can further improve model parsimony and obtain simpler interpretations by fitting latent class models that utilize the ordinality. One such model of this type also treats X as ordinal, and assumes a linear-by-linear association between each classification and X. Specifically, the model uses scores $\{u_i\}$ for the observed scale and scores $\{x_l\}$ for the latent classes, and has form

$$\log m_{hijl} = \mu + \lambda_h^A + \lambda_i^B + \lambda_j^C + \lambda_l^X + \beta^{AX} u_h x_l + \beta^{BX} u_i x_l + \beta^{CX} u_j x_l. \tag{4.1}$$

This model is a member of one type of latent class association model for ordinal variables considered by Agresti and Jørgensen in unpublished work, in which the scores may be fixed or parameters. When $\{x_l\}$ have equal spacing (e.g., $x_{l+1} - x_l = 1$ for all l), then (4.1) implies the adjacent-categories linear logit relationship

$$\log\left[\frac{\Pr(X = l + 1)}{\Pr(X = l)}\right] = \alpha_l + \beta^{AX} u_h + \beta^{BX} u_i + \beta^{CX} u_j, \quad l = 1, \ldots, L - 1 \tag{4.2}$$

for the effects of the observed variables on the latent classification.

It follows from Goodman (1985) that model (4.1) should fit well when there is an underlying multivariate normal distribution, with zero partial correlation between pairs of observed variables, given the latent variable. This model with $\beta^{AX} = \beta^{BX} = \beta^{CX}$ is a very parsimonious QLC model. When the scores are fixed, a single parameter (β) determines all odds ratios between observed variables and X. We refer to this special case as a *linear-by-linear quasi-symmetric latent class* ($L \times L$ QLC) *model.*

For the $L \times L$ QLC model, odc s ratios comparing marginal distributions take especially simple form when the single-factor parameters are linearly related. Namely, suppose

$$\lambda_i^r = \lambda_i + a_i \delta_r, \quad r = 1, \ldots, R \text{ and } i = 1, \ldots, I, \tag{4.3}$$

where $\{a_i\}$ are monotone increasing scores and $\{\lambda_i\}$ and $\{\delta_r\}$ satisfy constraints such as $\lambda_1 = \delta_1 = 0$. Then, $\lambda_i^A - \lambda_i^B = a_i(\delta_A - \delta_B)$, and when $\{a_{i+1} - a_i = 1\}$,

$$\frac{\Pr(A = i + 1 \mid X = l)/\Pr(A = i \mid X = l)}{\Pr(B = i + 1 \mid X = l)/\Pr(B = i \mid X = l)} = \exp(\delta_A - \delta_B) \tag{4.4}$$

for all i and l. The ratings distributions are then stochastically ordered, with $\{\delta_A > \delta_B\}$ equivalent to $\{\Pr(A > i \mid X) > \Pr(B > i \mid X) \text{ for } i = 1, \ldots, I - 1\}$.

5. Inference for QLC Models

To conduct inference about QLC models, we assume $\{n_{hij}\}$ in the observed A–B–C contingency table have independent Poisson(m_{hij}) distributions; or, equivalently for the parameters of interest, we condition on n and assume a multinomial(n, $\{\pi_{hij}\}$) distribution. One can fit the models using standard methods for latent class models (Goodman, 1974; Haberman, 1979), such as the EM algorithm. In that algorithm, the E (expectation) step used proportional fitting to approximate counts in the full A–B–C–X table using the observed A–B–C counts and the working conditional distribution of X, given the observed responses. In the M (maximization) step we treated those approximate counts as data in the standard iterative reweighted least squares algorithm for fitting loglinear models. We fitted the models of Sections 3 and 4 using the GLIM package, supplying appropriate macros to combine the E step with the ordinary GLIM fitting of Poisson loglinear models. The routine has slow convergence but is simple and seems insensitive to starting values, at least when the log-likelihood has a unique local maximum (such as usually occurs for small L). One can also fit the models using some existing programs for latent class models, such as LAT (Haberman, 1979) and NEWTON (Haberman, 1988). One can use fitted models to estimate probabilities $\{\pi_{+++l}\}$ of classification in various latent classes as well as conditional probabilities $\{\rho_{i|ri}\}$ of various responses, given the latent class.

One can obtain estimated standard errors for model parameter estimators by inverting the estimated information matrix for the nonlinear model for the observed table. Or, one can apply a general formula that Louis (1982) gave for estimating the observed information when using the EM algorithm. This formula provides an enlightening view of how the information for the model for the observed I^R table compares to that of the loglinear model for the $I^R \times L$ table that also treats the latent variable as observed. Let $\mathbf{Y}$ denote the counts in the I^R observed table, and let $\mathbf{Z}$ denote counts in the $I^R \times L$ cross-classification of $\mathbf{Y}$ with the latent variable. Louis showed that the observed information I_Y for the model for $\mathbf{Y}$ is related to the expected full-data observed information I_Z of $\mathbf{Z}$ by

$$I_Y = I_Z - I_{Z|Y},$$

where $I_{Z|Y}$ denotes the expected information for the conditional distribution of $\mathbf{Z}$ given $\mathbf{Y}$.

In the latent class model context, let $\mathbf{X}$ denote the model matrix for the Poisson loglinear model for the full $I^R \times L$ table; that is, the model has form $\log[\mathrm{E}(\mathbf{Z})] = \mathbf{X}\boldsymbol{\beta}$. Let $\mathbf{D}$ be a diagonal matrix with the elements of $\mathrm{E}(\mathbf{Z})$ on the main diagonal. Let $\mathbf{V}$ be a block-diagonal matrix, each block of which has multinomial covariance structure for cell counts across the latent dimension at a fixed level of the observed variables. For instance, when $R = 3$, each block is an $L \times L$ multinomial covariance matrix for $\{n_{hij1}, \ldots, n_{hijL}\}$ implied by the model, conditional on their sum n_{hij+}. Then,

$$I_Z = \mathbf{X'DX}, \quad I_{Z|Y} = \mathbf{X'VX}, \quad \text{and} \quad I_Y = \mathbf{X'(D - V)X}.$$

Thus, the covariance matrix for $\hat{\boldsymbol{\beta}}$ is approximately $[\mathbf{X'(D - V)X}]^{-1}$. When $L = 1$, all cell counts in the full table are observed, so $\mathbf{V} = \mathbf{0}$ and we obtain the usual information matrix for a Poisson loglinear model (e.g., Agresti, 1990, p. 179). When $L \geqslant 2$, there is a reduction in information from not observing the latent variable, and parameter estimators have larger variances in the resulting nonlinear model.

To test the fit of a model, one can use chi-squared goodness-of-fit statistics to compare $\{n_{hij}\}$ to model fitted values. The residual degrees of freedom (df) equal $I^R - L(RI - R + 1)$ for the ordinary latent class model, $I^R - [(R - 1)(I - 1) + LI]$ for the QLC model, and $I^R - (RI + L - R + 1)$ for the $L \times L$ QLC model. The QLC and $L \times L$ QLC models are special cases of the quasi-symmetry model for the I^R cross-classification of the observed variables, which has df $= I^R - (I - 1)(R - 1) - (I + R - 1)!/[(I - 1)!R!]$. When $L = (I + R - 1)!/(R!\ I!)$, the QLC model is equivalent to quasi-symmetry for the I^R table. When L exceeds this value, the QLC model is unidentifiable. The QLC models, being highly parsimonious, are applicable to more situations than ordinary latent class models. When $I = 2$, for instance, unlike the ordinary model, the QLC model is unsaturated when $\{R = 3; L = 2\}$.

6. Application to Modeling Interrater Agreement

The modeling of multirater agreement is an application in which the assumption of local independence is reasonable. Suppose R raters rate the same sample of subjects on a categorical scale, such as (positive, negative) for diagnosis of whether subjects have a certain disease. When ratings are done "blindly," ratings of a given subject by different raters are independent. If subjects having "true" rating in a given category are relatively homogeneous, then ratings by different raters may be nearly independent within a given true rating class. For instance, when $I = 2$, the agreement structure specified by the 2^R joint distribution for the R ratings may be a mixture of two distributions, statistical independence among raters for subjects whose true rating is positive, and statistical independence among raters for subjects whose true rating is negative. A QLC model is then appropriate if there is no three-factor interaction among rater, response, and subject. The "true" rating scale or the scale generating homogeneous subsets of subjects need not have the same categories as the observed scale, so L need not equal I.

Interrater agreement has two components—distinguishability of categories and lack of bias. For subjectively defined categorical scales, distinguishability refers to how well an expert rater can distinguish between pairs of categories. For two raters, distinguishability increases as the association in their joint distribution becomes more strongly positive, in the sense that odds ratios of the type $[\Pr(A = i, B = i)\Pr(A = j, B = j)/\Pr(A = i, B = j)\Pr(A = j, B = i)]$ become larger (Darroch and McCloud, 1986). Bias decreases as their marginal distributions become more nearly equivalent. Strong agreement, in terms of relatively high probability of identical ratings, requires both similar marginal distributions and strong positive association. In QLC models, variation in marginal distributions is addressed by variation in the $\{\delta_r\}$ parameters [see (3.4)], and strength of association is induced by the common association between each observed variable and the latent variable. For instance, when $I = L = 2$ in the QLC model, the marginal odds ratio between each pair of observed variables is monotone increasing in λ for $\lambda > 0$ (for fixed single-factor parameters), equaling 1 when $\lambda = 0$. The strength of agreement improves in the two-way tables relating pairs of raters as $\{\delta_r\}$ move toward uniformity and the association between each rater and X increases.

We now use Table 1 to illustrate quasi-symmetric latent class models. The original classifications were made on a five-point scale, but for simplicity Landis and Koch (1977) and Schouten (1982) analyzed the data using the binary representation in Table 1, whereby category 2 combines the third, fourth, and fifth categories from the five-point scale. Even with this simplification, the data are sparse. Table 2 reports likelihood-ratio statistics for testing the fit to these data of several latent class models. Because of the sparseness, we use these statistics primarily for comparing models with a fixed number of latent classes.

The latent class model with $L = 1$ latent class is simply the model of mutual independence of the seven ratings. It fits poorly, as one would expect. For $L = 2$, the QLC model and the ordinary latent class (LC) model have substantial lack of fit. For instance, they give fitted counts of about 23, compared to the observed 34, for the cell corresponding to a rating of 1 by each rater. Models having $L = 3$ fit much better than those having $L = 2$. In addition, the parsimonious quasi-symmetric models (i.e., the QLC and $L \times L$ QLC models) fitted essentially as well as the general quasi-symmetry model for the 2^7 table, so it is unnecessary to consider $L = 4$ latent classes. In fact, the QLC model with $L = 4$ is identical to the general quasi-symmetry model. Further discussion refers only to models having $L = 3$.

When $I = 2$ and $L = 3$, the ordinary latent class model is equivalent to the generalization of model (4.1) in which $\{\lambda_l = \beta x_l\}$ are parameters, and different $\{\lambda_l\}$ apply in each association. For the fit of that model, the estimated $\{\beta x_l\}$ are monotone increasing in l for all raters except B. The QLC model with $I = 2$ and $L = 3$ is equivalent to the homogeneous version of model (4.1) in which $\{\beta x_l\}$ are parameters and are identical for each association. Table 2 shows that this model is much more parsimonious than the ordinary LC model, yet does not give a significantly poorer fit. Expressing each association in the form $\beta u_i x_l$ and setting $u_2 - u_1 = 1$ and $x_1 = 0$ for identifiability, we obtain ML estimates of $\{\beta x_l, l = 1, 2, 3\}$ for the QLC model of 0, 4.65, and 8.88. The estimated log odds ratio between each rater and X is 4.65 for the first two levels of X, and 4.23 for the last two levels of X. These two estimates suggest using the simpler model in which they are identical, which corresponds to setting $\{x_l\} = \{0, 1, 2\}$. Table 2 shows that this model, the $L \times L$ QLC model, also fits well. It yields simple interpretations, which we discuss next.

In the $L \times L$ QLC model, $\hat{\beta} = 4.38$. Thus, the estimated odds ratio between each observed variable and levels x_a and x_b of X is $\exp[4.38|x_a - x_b|]$. For instance, the odds that a rater selects category 1 are estimated to be $\exp[4.38(2)] = 6{,}374$ times higher for subjects in the first latent category than for subjects in the third category. From the inverse of the estimated information matrix, the estimated standard error for $\hat{\beta}$ is .374. An approximate 95% confidence interval for the odds ratio just described is $\exp\{2[4.38 \pm 1.96(.374)]\}$, or $(1.5, 27.8) \times 10^3$. The estimated standard error of $\hat{\beta}$ using Louis's (1982) estimator of the information matrix is .422, leading to an approximate 95% confidence interval of $(1.2, 33.7) \times 10^3$. The intervals are crude, using standard error and normal sampling distribution approximations that may be poor for such sparse data for a nonlinear model. However, the intervals make clear that there is very strong association between each rating and the latent rating, with the

Table 2
Likelihood-ratio statistics for testing fit of latent class models to Table 1

No. latent classes	Model	Likelihood-ratio statistic	Degrees of freedom
1	Independence	476.8	120
	Quasi-symmetry	23.7	114
2	Ordinary LC	62.4	112
2	QLC	67.6	118
3	Ordinary LC	15.3	104
3	QLC	27.5	116
3	$L \times L$ QLC	27.7	117
3	$L \times L$ QLC + Marginal homogeneity	259.4	123

point estimate of the strength being quite imprecise. The corresponding fitted odds ratio estimates between pairs of raters are also strong, varying between 7.2 and 394.2.

Table 2 shows that the simpler $L \times L$ QLC model assuming marginal homogeneity (i.e., $\delta_A = \cdots = \delta_G$) fits poorly. For the $L \times L$ QLC model, Table 3 shows the estimated $\{\delta_r\}$ and estimated standard errors of the estimates of $\{\delta_r - \delta_t\}$, calculated using the inverse of the estimated information matrix. The estimated $\{\delta_r\}$ (which we scaled to satisfy $\delta_A = 0$) show that rater B tends to make the greatest number of ratings of carcinoma, and D and F the least. For instance, $\hat{\delta}_B - \hat{\delta}_F = 7.07$ means that in each latent class, the estimated odds of a diagnosis of carcinoma are exp(7.07) = 1,180 times higher for B than for F. Based on the estimated standard errors of the differences, a 90% simultaneous Bonferroni comparison of the 21 pairs shows that the marginal differences for the (A, E), (A, G), (B, E), (C, D), (D, F), and (E, G) pairs of raters are not statistically significant. Figure 1 shows results of such a comparison.

Table 4 reports for each rater the estimated probability of carcinoma diagnosis, conditional on the latent class. Note that odds estimates using these values yield the $\{\exp(\delta_r - \delta_t)\}$ estimates just described; for instance, $\exp(\hat{\delta}_B - \hat{\delta}_A) = (.148/.852)/(.021/.979)$, and similarly for the other latent classes. The fitted probabilities suggest an interpretation for the latent classes. The first latent class consists of cases that all raters (except occasionally B) agree show no carcinoma. The third latent class consists of cases in which raters A, B, E, and G agree there is carcinoma, and C and D usually agree. The second class consists of the cases of strong disagreement, whereby C, D, and F rarely diagnose carcinoma, but A, B, E, and G usually do. The estimated proportions in the three latent classes are .37, .20, and .43, so about 20% of the cases are in the problematic class.

In summary, the $L \times L$ QLC model with $L = 3$ has only three more parameters than the mutual independence model, yet it fits well and provides simple interpretations. It requires one parameter to describe associations between each rater and the latent variable, and six parameters to describe variation in the marginal distributions of the raters. These seven parameters lend insight in describing the structure of agreement in a table having 128 cells. There is very strong but uniform positive association between each rating and the latent rating, which induces strong association between pairs of ratings; however, there is substantial marginal heterogeneity among the ratings, which causes

Table 3

Estimated $\{\delta_r\}$ and $\{\delta_r - \delta_t\}$ (with estimated standard errors in parentheses), for linear-by-linear quasi-symmetric latent class model with L = 3 classes

		Pathologist						
		A	B	C	D	E	F	G
δ_r:		.00	2.11	−3.16	−4.42	.80	−4.96	.00
$\delta_r - \delta_t$:	A		−2.11 (.61)	3.16 (.65)	4.42 (.66)	−.80 (.58)	4.96 (.66)	.00 (.55)
	B			5.27 (.84)	6.53 (.86)	1.32 (.60)	7.07 (.86)	2.11 (.61)
	C				1.26 (.47)	−3.96 (.72)	1.80 (.48)	−3.16 (.65)
	D					−5.22 (.73)	.54 (.40)	−4.42 (.66)
	E						5.76 (.74)	.80 (.58)
	F							−4.96 (.66)

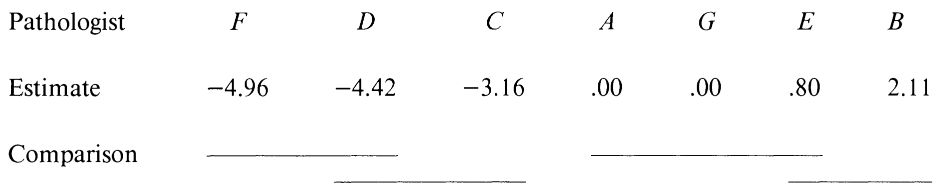

Figure 1. Result of 90% Bonferroni simultaneous comparison of $\{\hat{\delta}_r\}$.

Table 4
Estimated probabilities of diagnosing carcinoma, for linear-by-linear quasi-symmetric latent class model with $L = 3$ classes

	Pathologist						
	A	*B*	*C*	*D*	*E*	*F*	*G*
Pr(carc. | $X = 1$)	.021	.148	.001	.000	.044	.000	.021
Pr(carc. | $X = 2$)	.627	.933	.067	.020	.789	.012	.627
Pr(carc. | $X = 3$)	.993	.999	.852	.619	.997	.485	.993

heterogeneity in pairwise levels of agreement. Whatever lack of agreement exists seems due more to bias than to category indistinguishability. If the raters could calibrate themselves to achieve marginal homogeneity, then this model would simplify to complete symmetry in the joint ratings table, and hence uniformity in the pairwise agreement structure.

7. Comments

For the case $I = 2$, the models discussed in this article are related to special cases of the Rasch model (Rasch, 1961). That model, for R items and S subjects, has the form

$$\log(\phi_{sr1}/\phi_{sr2}) = \alpha_s + \delta_r.$$

That is, it assumes no three-factor interaction for $\{\phi_{sri}\}$, with a binary response. It follows from Tjur (1982) that one can obtain conditional ML estimates of $\{\delta_r\}$ in the Rasch model by fitting the quasi-symmetry model to the 2^R table that cross-classifies responses on the R items [e.g., model (2.3) for $R = 3$]. For Table 1, the conditional ML estimates (scaled so $\delta_A = 0$) are {.00, 2.07, −3.39, −4.71, .80, −5.38, .00}. These are similar to the estimates reported in Table 3 of the analogous parameters in the simpler $L \times L$ QLC model, which is not surprising since that model also fit well. When $I = 2$, the QLC models correspond to a class of latent class models introduced by Lindsay et al. (1991), which they referred to as Rasch mixture models. In that case, the QLC model also corresponds to a logistic latent class model presented by Uebersax (1993) for rater agreement, having equal measurement error rates across raters. Andrich (1978), Clogg (1988), and Rost (1988) considered other latent class approaches that have similarities with models discussed in this article.

When quasi-symmetric latent class models hold, they provide the advantage of simple interpretation. However, they are so simple that they may have limited scope. In some applications, lack of fit may occur because local independence holds at the subject level but does not hold for a latent variable having few latent classes. For instance, subject homogeneity may be determined by a continuous variable, in which case homogeneity within levels of X may occur to a decent approximation only for relatively large L. Or, lack of fit may occur because of violations of the assumption of no three-factor interaction for $\{\phi_{sri}\}$. From results in Lindsay et al. (1991), it follows that one can check this assumption by comparing the fit of the QLC model to that of the ordinary LC model having the same number of latent classes. Though the scope of QLC models may be limited, we believe they are worthy of notice because of the economical description available when they do fit well.

Acknowledgements

We thank Bent Jørgensen for suggesting the usefulness of latent class approaches to modeling agreement, and John Uebersax for several helpful comments. This research was partially supported by National Institutes of Health Grant GM 43824.

Résumé

Supposons que nous observons des réponses à plusieurs variables catégorielles ayant la même échelle. Nous considérons les modèles à classe latente dont la classification conjointe respecte la condition de quasi-symétrie. Ces modèles s'appliquent lorsque les distributions des réponses spécifiques au sujet sont telles que (i) les réponses aux différentes variables sont indépendantes pour un sujet donné et (ii) que les odds ratios comparant les distributions marginales des variables sont identiques pour chaque sujet. Ces hypothèses sont souvent raisonnables dans la modélisation de l'accord entre plusieurs évaluateurs lorsque l'échantillon de sujets est évalué indépendamment par différents observateurs. Dans cette application, les paramètres du modèle décrivent deux composantes de l'accord entre observateurs, l'intensité de l'association entre les classifications par paire d'observateurs et le degré d'hétérogénéité parmi les distributions marginales des observateurs. Nous illustrons ces modèles en analysant un fichier de données pour lequel plusieurs pathologistes ont classifié 118 patients en terme de présence–absence de carcinome, conduisant à plusieurs classifications catégorielles avec la même échelle binaire. Un modèle correctement ajusté aux données possède une classification latente qui différentie les sujets pour lesquels il y a concordance de ceux pour lesquels il y a discordance de jugement.

REFERENCES

Agresti, A. (1988). A model for agreement between ratings on an ordinal scale. *Biometrics* **44,** 539-548.

Agresti, A. (1990). *Categorical Data Analysis.* New York: Wiley.

Aickin, M. (1990). Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. *Biometrics* **46,** 293–302.

Aitkin, M., Anderson, D., and Hinde, J. (1981). Statistical modelling of data on teaching styles (with Discussion). *Journal of the Royal Statistical Society, Series A* **144,** 419–461.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika* **43,** 561-573.

Becker, M. (1990). Quasisymmetric models for the analysis of square contingency tables. *Journal of the Royal Statistical Society, Series B* **52,** 369–378.

Becker, M. and Agresti, A. (1992). Loglinear modeling of pairwise interobserver agreement on a categorical scale. *Statistics in Medicine* **11,** 101–114.

Caussinus, H. (1965). Contribution à l'analyse statistique des tableaux de correlation. *Annales de la Faculté des Sciences de l'Université de Toulouse* **29,** 77–182.

Clogg, C. C. (1979). Some latent structure models for the analysis of Likert-type data. *Social Science Research* **8,** 287–301.

Clogg, C. C. (1981). Latent structure models of mobility. *American Journal of Sociology* **86,** 836–868.

Clogg, C. C. (1988). Latent class models for measuring. In *Latent Trait and Latent Class Models,* R. Langeheine and J. Rost (eds), 173–205. New York: Plenum Press.

Darroch, J. N. and McCloud, P. I. (1986). Category distinguishability and observer agreement. *Australian Journal of Statistics* **28,** 371–388.

Dillon, W. R. and Mulani, N. (1984). A probabilistic latent class model for assessing inter-judge reliability. *Multivariate Behavioral Research* **19,** 438–458.

Espeland, M. A. and Handelman, S. L. (1989). Using latent class models to characterize and assess relative error in discrete measurements. *Biometrics* **45,** 587–599.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61,** 215–231.

Goodman, L. A. (1985). 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.

Haberman, S. J. (1979). *Analysis of Qualitative Data, Vol. 2.* New York: Academic Press.

Haberman, S. J. (1988). A stabilized Newton–Raphson algorithm for log-linear models for frequency tables derived by indirect observation. *Sociological Methodology* **18,** 193–211.

Landis, J. R. and Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* **33,** 363–374.

Lindsay, B., Clogg, C. C., and Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association* **86,** 96–107.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* **44,** 226–233.

McCutcheon, A. L. (1987). *Latent Class Analysis.* Beverly Hills, California: Sage Publications.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, Vol. 4,* J. Neyman (ed.), 321–333. Berkeley, California: University of California Press.

Rost, J. (1988). Rating scale analysis with latent class models. *Psychometrika* **53,** 327–348.

Schouten, H. J. A. (1982). Measuring pairwise interobserver agreement when all subjects are judged by the same observers. *Statistica Neerlandica* **36,** 45–61.

Tjur, T. (1982). A connection between Rasch's item analysis model and a multiplicative Poisson model. *Scandinavian Journal of Statistics* **9,** 23–30.

Uebersax, J. S. (1993). Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of the American Statistical Association* **88,** in press.

Uebersax, J. S. and Grove, W. M. (1990). Latent class analysis of diagnostic agreement. *Statistics in Medicine* **9,** 559–572.

Received January 1991; *revised August* 1991; *accepted August* 1991.