Binormal Association-Marginal Models for ROC Analysis based on Correlated Ordinal Rating Data

Joseph B. Lang and Thor Aspelund

Department of Statistics and Actuarial Science
University of Iowa, IA 52242 USA

Technical Report #299

November 2000

Binormal Association-Marginal Models for ROC Analysis based on Correlated Ordinal Rating Data[*]

Joseph B. Lang[†] and Thor Aspelund[‡]

November 7, 2000

Abstract

Correlated ordinal response data are used to estimate and compare two receiver operating characteristic (ROC) curves. The data are modeled using a flexible, new class of binormal association-marginal (BAM) models. BAM models use the latent binormal structure of classic signal detection theory to model each ordinal response marginal distribution. In contrast to bivariate binormal models, BAM models do not impose the added restriction that the ordinal responses have joint distributions that are determined by latent bivariate normal distributions. Instead, the association structure of the ordinal variables is directly specified using standard loglinear models. The maximum likelihood fitting program BAMROC, which uses an algorithm related to those used to fit composite-link generalized linear marginal models, is described. The method is illustrated through the analysis of a neonatal radiograph data set and a small simulation study.

KEYWORDS: bivariate binormal model, diagnostic performance, loglinear model, marginal model, maximum likelihood, multivariate ordinal response, ROC curve, signal detection theory.

1 Introduction

Receiver operating characteristic (ROC) analysis is commonly used to evaluate the performance of signal detection procedures (modalities), such as medical imaging or screening procedures. For convenience, we use medical imaging as our canonical example, and adopt the corresponding language. These modalities are designed to aid in the diagnosis of units (e.g. patients), which, in theory, can be classified as "case" or "control." Often, two or more modalities are to be

[*] Supported in part by National Institutes of Health research grant RO1 CA 62362.

[†] Dept. of Statistics and Actuarial Science, Univ. of Iowa, IA 52242, email: jblang@stat.uiowa.edu

[‡] Dept. of Statistics and Actuarial Science, Univ. of Iowa, IA 52242, email: aspelund@stat.uiowa.edu

compared vis-a-vis their diagnostic performance. This paper outlines a method that can be used to empirically investigate and compare the diagnostic performances of two or more modalities.

Consider a situation whereby a "reader" uses two modalities to evaluate a sample of n_1 control and n_2 case units; the case-control status is unknown to the reader. Each unit is assigned two suspicion rating scores, one for each modality. Here, we assume that the same scale, 1 to R (≥ 2), where $1 =$ "definitely not a case" and $R =$ "definitely a case," is used for both modalities.

The resulting suspicion rating data can be displayed using two $R \times R$ cross-classification tables, one for the controls and one for the cases. The (i, j) cell in the case table contains the number of cases rated i using the first modality and j using the second modality. The row [column] margins of the two tables could be used to estimate measures of the diagnostic performance of modality 1 [2]. To compare the two estimates, however, it is important to note that the row and column marginal counts are *not* generally independent because the same sample of units was evaluated using both modalities. This is the "correlated" data case considered in Metz et al. (1984). When each modality is used to evaluate different samples, the comparison is more straightforward (Metz and Kronman, 1984). Drawbacks to this independent sample design include (i) decreased power for discerning modality differences and (ii) the association between the two modality ratings cannot be described.

The paper is organized as follows. Section 2 introduces a paired-comparison study of neonatal radiograph evaluation methods (Franken et al. 1992), and gives some of the data. Section 3 describes the classic model of signal detection theory and the corresponding ROC curves. Binormal association-marginal (BAM) models are introduced in Section 4. Section 5 briefly discusses currently available methods and models. In particular, the closely related bivariate binormal model of Metz et al. (1984) is described. We argue that the flexible class of BAM models serve as an attractive alternative to the bivariate binormal model. Section 6 introduces the maximum likelihood fitting algorithm used in the computer program BAMROC; the algorithm is related to those used to fit composite-link generalized linear marginal models (cf. Glonek and McCullagh 1995 or Glonek 1996). Maximum likelihood estimation of areas under ROC curves and the differences between them is also discussed. Section 7 provides an analysis of the neonatal radiograph data using the methods discussed in the paper. To illustrate model differences, simulated tables are analysed in Section 8. Section 9 provides a brief summary and discussion.

2 Example: Neonatal Radiograph Data

A paired-comparison study of neonatal radiographs was discussed in Franken et al. (1992). Although four radiologists participated in the study, we restrict attention to one of them in this paper. The radiologist used a 5-point suspicion scale to rate 33 normal (control) and 67 abnormal (case) radiographs. Each of the 100 radiographs was viewed twice, once using a video image (modality 1) and once using a plain film image (modality 2). It is of interest to determine which image, video or plain film, is a better diagnostic. The rating data are given in Table 1 below.

Table 1. Neonatal Radiograph Rating Data

		Plain Film					
		1	2	3	4	5	
	1	4	1	1	0	0	6
	2	4	8	3	1	1	17
Video	3	0	2	2	0	0	4
	4	0	3	1	1	0	5
	5	0	0	0	1	0	1
		8	14	7	3	1	33

Normals

		Plain Film					
		1	2	3	4	5	
	1	1	0	1	2	0	4
	2	1	2	1	1	0	5
Video	3	0	2	1	1	1	5
	4	0	2	3	4	6	15
	5	1	1	3	6	27	38
		3	7	9	14	34	67

Abnormals

To get a rough idea of the diagnostic performance of video imaging, we can compare the two row marginal empirical distributions. The lowest suspicion score '1' was assigned to 6 of 33 normals and only 4 of 67 abnormals. Similarly, only 1 of 33 normals was assigned the highest suspicion score '5,' while 38 of the 67 abnormals were assigned a '5'. Evidently, there is some diagnostic capabilities of video imaging. In the same way, we can roughly assess the diagnostic performance of plain film imaging by comparing the two column marginal empirical distributions.

3 Measuring Diagnostic Performance using ROC Curves

This section describes a commonly-used, formal approach for measuring and comparing diagnostic performances of modalities. Before describing the approach, we make an observation regarding the comparison of diagnostic performances of the two modalities used in the example of the previous section.

A comparison of diagnostic performances of video and plain film imaging arguably should *not* be based on the direct comparison of row and column marginal empirical distributions of Table 1. Unless the rating scales for the two modalities are used in exactly the same way, a comparison like "6 of 33 normals using video imaging compared to 8 of 33 normals using plain film imaging were rated '1'" is not very meaningful; the rating '1' could mean something very different for the two modalities. Often, the scientific objective is to measure diagnostic performance in a way that does not depend on the way the rating scale is used.

In contrast to non-parametric measures of diagnostic performance (cf. Delong et al. 1988), the classic model of signal detection theory, which is described below, affords a measure of diagnostic performance that is independent of the way the rating scale is used. We point out that with this classic parametric model, the measured diagnostic performances of two modalities can be identical even when the ordinal response marginal distributions corresponding to the two modalities are quite different.

The classic model of signal detection theory for ordinal rating data was originally introduced in the psychometric literature (e.g., see Green and Swets 1966, Dorfman and Alf 1969), and has been used in many other disciplines since then. Using the language of the current paper, the model assumes that each unit "generates" an imperfectly-observable (latent) suspicion score. The latent scores for controls are assumed to be realizations of a continuous random variable X^*, and the latent scores for cases are realizations of a continuous random variable Y^*. The model states that the observable (or manifest) ordinal response variables, say X and Y, have distributions that are determined according to $P(X \leq j) = P(X^* \leq c_j)$ and $P(Y \leq j) = P(Y^* \leq c_j)$. That is, the manifest responses are discretized versions of the latent suspicion scores; the cutpoints c_j's are independent of case-control status. Diagnostic performance (i.e. signal detection capability) measures are based on some comparison of the latent X^* and Y^* distributions. As an example, if the modality is an effective method of diagnosis, we might expect that Y^* is stochastically larger than X^*.

A simple way to graphically compare the X^* and Y^* distributions is to create a parametric curve of $P(X^* > c)$ by $P(Y^* > c)$, where the cutpoint or threshold c is the parameter and runs from $-\infty$ to ∞. This is the (latent) receiver operating characteristic (ROC) curve (cf. Metz 1978, Hanley and McNeil 1982), which gives a description of diagnostic performance that is independent of the way the manifest rating scale is used. The probabilities plotted in an ROC curve have particularly nice interpretations. If the decision rule is to classify as "case" if the (latent) suspicion score exceeds the threshold value c, then $P(X^* > c)$ is the false positive probability and $P(Y^* > c)$ is the true positive probability.

Functionals of the latent ROC curve can be used as diagnostic performance measures. One commonly used measure is the "area under the curve" (AUC) or area index (cf. Hanley 1998). This area can be shown to equal $P(Y^* > X^*)$ and, hence, has a nice interpretation as the probability that the latent suspicion score for a randomly selected case is higher than for a randomly selected control. Two or more modalities can be compared by measuring the differences between the AUC's.

Note that we do not use a manifest measure like $P(Y > X)$ as a measure of diagnostic performance, because this measure is not manifest scale independent. Of historical interest, Bamber (1975) referred to measures based on the manifest and latent variables as diagnostic performance (not to be confused with our more general use of the word 'performance') and diagnostic capacity, respectively. Using Bamber's terminology, non-parametric approaches estimate diagnostic performance and our approach estimates diagnostic capacity. It is important to understand that these two measures are different entities which can be very different numerically.

4 The Binormal Association-Marginal Model

Let X_m and Y_m represent the observable (or manifest) suspicion scores for a randomly selected control and a randomly selected case, respectively, when modality m is used. These manifest ordinal variables take on values $1, \ldots, R$, and are used to create the cross-classification tables discussed above. The ultimate goal is to use data which are realizations of the manifest variables to estimate and compare the latent ROC curves.

4.1 Random component

Let N_{1ij} = the number of controls for which $(X_1, X_2) = (i, j)$ and let N_{2ij} = the number of cases for which $(Y_1, Y_2) = (i, j)$. It follows that

$$\{N_{kij} : i, j = 1, \ldots, R\} \text{ indep } \sim \ Mult(n_k, \{\pi_{kij} : i, j = 1, \ldots, R\}), \quad k = 1, 2.$$

where $\pi_{1ij} = P(X_1 = i, X_2 = j)$ and $\pi_{2ij} = P(Y_1 = i, Y_2 = j)$. Assume that $n_{kij} \leftarrow N_{kij}$ are the observed counts.

4.2 Systematic component

The systematic component for a binormal AM model comprises an association and a marginal component.

Marginal Model: The marginal model implies that the manifest ordinal variables have distributions that are determined by latent continuous variables,

$$\begin{aligned}
\pi_{1i+} &= P(X_1 = i) &&= P(c_{1,i-1} < X_1^* \leq c_{1,i}) \\
\pi_{2i+} &= P(Y_1 = i) &&= P(c_{1,i-1} < Y_1^* \leq c_{1,i}) \\
\pi_{1+i} &= P(X_2 = i) &&= P(c_{2,i-1} < X_2^* \leq c_{2,i}) \\
\pi_{2+i} &= P(Y_2 = i) &&= P(c_{2,i-1} < Y_2^* \leq c_{2,i}), \quad i = 1, \ldots, R,
\end{aligned}$$

where $-\infty \equiv c_{m,0} < c_{m,1} < \cdots < c_{m,R} \equiv \infty, \ m = 1, 2.$

A binormal model (cf. Green and Swets 1966, Dorfman and Alf, Jr. 1969) for modality m is specified by assuming, without loss of generality, that X_m^* and Y_m^* have normal distributions with means 0 and μ_m and variances 1 and $\sigma_m^2 \equiv \exp(2\xi_m)$, respectively. Although other distributions such as the gamma (cf. Dorfman et al. 1997) could be considered, we will restrict attention to the normal distribution in this paper.

In matrix notation, the marginal model for the k^{th} population can be specified as

$$\mathbf{M}_k \boldsymbol{\pi}_k = \mathbf{f}_k(\boldsymbol{\beta}),$$

where $\mathbf{M}_k\boldsymbol{\pi}_k$ is the $2(R-1)\times 1$ vector of non-redundant row and column marginal cumulative probabilities and $\mathrm{f}_k(\boldsymbol{\beta})$ is a vector of normal probabilities which depend on $\boldsymbol{\beta}$, the vector that contains cutpoint, mean, and variance parameters. As an example, if suspicion scores are recorded on an $R=5$ level ordinal scale and $k=1$ corresponds to the control population, then

$$\mathbf{M}_1\boldsymbol{\pi}_1 \equiv \begin{bmatrix} \pi_{11+} \\ \pi_{11+}+\pi_{12+} \\ \vdots \\ \pi_{11+}+\pi_{12+}+\pi_{13+}+\pi_{14+} \\ \pi_{1+1} \\ \vdots \\ \pi_{1+1}+\pi_{1+2}+\pi_{1+3}+\pi_{1+4} \end{bmatrix} = \begin{bmatrix} \Phi(c_{11}) \\ \Phi(c_{12}) \\ \vdots \\ \Phi(c_{14}) \\ \Phi(c_{21}) \\ \vdots \\ \Phi(c_{24}) \end{bmatrix} \equiv \mathrm{f}_1(\boldsymbol{\beta}),$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Similarly, assuming that $k=2$ corresponds to the case population,

$$\mathbf{M}_2\boldsymbol{\pi}_2 \equiv \begin{bmatrix} \pi_{21+} \\ \pi_{21+}+\pi_{22+} \\ \vdots \\ \pi_{21+}+\pi_{22+}+\pi_{23+}+\pi_{24+} \\ \pi_{2+1} \\ \vdots \\ \pi_{2+1}+\pi_{2+2}+\pi_{2+3}+\pi_{2+4} \end{bmatrix} = \begin{bmatrix} \Phi((c_{11}-\mu_1)/\exp(\xi_1)) \\ \Phi((c_{12}-\mu_1)/\exp(\xi_1)) \\ \vdots \\ \Phi((c_{14}-\mu_1)/\exp(\xi_1)) \\ \Phi((c_{21}-\mu_2)/\exp(\xi_2)) \\ \vdots \\ \Phi((c_{24}-\mu_2)/\exp(\xi_2)) \end{bmatrix} \equiv \mathrm{f}_2(\boldsymbol{\beta}).$$

Association Model: In this paper, we restrict attention to loglinear association models specified as $\log\boldsymbol{\pi} = \mathbf{X}\boldsymbol{\tau}$. To avoid redundancies and contradictions with constraints imposed by the marginal model, it is assumed that $\mathbf{X}$ includes columns corresponding to population-specific intercepts (τ_k) and main-effects parameters ($\tau_{ki}^{(1)}, \tau_{kj}^{(2)}$); this loglinear model imposes no constraints on the marginal distributions. Some examples include,

$$\begin{array}{lll} A_I: & \log\pi_{kij} = \tau_k + \tau_{ki}^{(1)} + \tau_{kj}^{(2)} & \text{(independence)} \\ A_Q: & \log\pi_{kij} = \tau_k + \tau_{ki}^{(1)} + \tau_{kj}^{(2)} + \lambda_k I(i=j) & \text{(quasi−independence)} \\ A_L: & \log\pi_{kij} = \tau_k + \tau_{ki}^{(1)} + \tau_{kj}^{(2)} + \lambda\cdot i\cdot j & \text{(homogeneous linear−by−linear)} \\ A_S: & \log\pi_{kij} = \tau_k + \tau_{ki}^{(1)} + \tau_{kj}^{(2)} + \lambda_{kij} & \text{(saturated)} \end{array}$$

Summarizing, the binormal association-marginal (BAM) model constrains the ordinal suspicion score marginal distributions by relating them to distributions of latent normal suspicion scores. The association between the pairs of dependent latent suspicion scores is not explicitly specified. Instead, the association between the two manifest ordinal suspicion scores is modeled directly using standard loglinear association models. Evidently, BAM models do not require explicit specification of the joint distribution of the latent normal suspicion scores. Because of this, the class of BAM models is very broad and non-restrictive. In addition, the BAM modeling approach effectively separates the tasks of specifying the marginal and association components of the model. This leads to simplified model interpretations; in particular, the association structures are directly interpretable on the manifest-response scale rather than on the latent-response scale.

5 A Comparison to Previous Methods and Models

Several methods and models have been proposed to compare two (or more) ROC curves using correlated data. Beam (1998) gives a nice description of many of the available methods. There are non-parametric methods based on multivariate U-statistic theory (Delong et al. 1988, Obuchowski 1997). Here, the objects of analysis are the ROC curves that are based directly on the manifest ordinal suspicion scores. This is in contrast to the latent-response-based ROC curves of the classic model of signal detection. As previously noted (see Section 3), whereas the latent ROC curves are invariant to the ordinal variable category definitions, the ROC curves based on the manifest ordinal suspicion scores are not. That is, changing the instructions on how readers are to use the ordinal suspicion scale will change the non-parametric ROC curve; it will not, in theory, change the latent-response-based parametric ROC curve. Bayesian hierarchical models that include cluster-specific random effects have been proposed (cf. Ishwaran and Gatsonis 2000). These models indirectly induce correlation between the ordinal suspicion scores through the introduction of random effect variables. Random-effects normal linear models for correlated functionals have been proposed (Thompson and Zucchini 1989, Dorfman et al. 1992). These methods do not attempt to model the original ordinal suspicion data, instead they focus attention on some summary measures, such as the areas under the curves. Fixed-effects, non-likelihood based parametric models fitted using generalized estimating equations (Toledano and Gatsonis 1995, 1996) or marginal likelihoods with correlation adjustment (Hanley and McNeil 1983) have also been proposed. These more ad-hoc methods can be very useful when there are many marginal responses, a situation when maximum likelihood fitting is difficult. One drawback to these approaches is that the goodness of fit of the association model cannot be directly assessed. The fixed-effects, likelihood based parametric model of Metz et al. (1984) is another alternative; it is most closely related to the method proposed herein.

Metz et al. (1984) introduced a bivariate generalization of the binormal model. Specifically, using the notation of this paper, Metz et al. assumed that

$$\begin{aligned}\pi_{1ij} &= P(c_{1,i-1} \leq X_1^* \leq c_{1i}, c_{2,j-1} \leq X_2^* \leq c_{2j}),\\ \pi_{2ij} &= P(c_{1,i-1} \leq Y_1^* \leq c_{1i}, c_{2,j-1} \leq Y_2^* \leq c_{2j}),\end{aligned}$$

where the cutpoints c_{mi} are defined as above. The bivariate binormal model specification is completed by assuming that the random vectors (X_1^*, X_2^*) and (Y_1^*, Y_2^*) have bivariate normal distributions. Metz et al. (1984) describes a maximum likelihood algorithm for fitting the model.

The bivariate binormal model restricts the marginal probabilities (the π_{ki+}'s and π_{k+i}'s) in exactly the same way as the binormal association-marginal model. This follows because the marginal distributions of the bivariate normal are normal. The obvious difference is in the way the association structure is modeled. The bivariate binormal model, owing to the properties of the bivariate normal distribution, necessarily implies a restrictive form of association which is measured in terms of correlation between the latent suspicion scores. In contrast, the BAM model association structures are (i) very flexible, ranging from very parsimonious to completely unrestricted, and (ii) measured directly on the manifest ordinal suspicion scores by means of standard categorical association models. It follows that BAM models are particularly useful when it is also important to model and describe the association structure between the ordinal suspicion scores.

There are other important differences between the bivariate binormal model and the BAM model. Because the association and marginal structures are considered separately, the BAM model approach allows for a direct test of goodness of fit of the binormal marginal model; this is not the case for the bivariate binormal model, as the association and marginal structures are not separable. We also point out that specification and maximum likelihood fitting of the generalized multivariate binormal model for three or more modalities becomes very unwieldy. In contrast, model specification and model fitting for BAM models is straightforward when there are three or more modalities.

6 ML Fitting using BAMROC

To facilitate simpler ML fitting by exploiting independence of case and control units, we write the binormal AM model as

$$\mathbf{L}_k^*(\boldsymbol{\pi}_k) \equiv \begin{bmatrix} \mathbf{1}^T\boldsymbol{\pi}_k \\ \mathbf{M}_k\boldsymbol{\pi}_k \\ \log\boldsymbol{\pi}_k \end{bmatrix} = \begin{bmatrix} 1 \\ \mathbf{f}_k(\boldsymbol{\beta}) \\ \tau_k\mathbf{1} + \mathbf{W}_k\boldsymbol{\omega}_k \end{bmatrix}, \quad \text{and} \quad \mathbf{A}_k\boldsymbol{\omega}_k = \mathbf{g}_k(\boldsymbol{\alpha}), \quad k = 1, 2, \tag{6.1}$$

where $\mathbf{A}_k\boldsymbol{\omega}_k$ is the vector of two-way association parameters only; it does not include the intercept or main effects. As an example, the homogeneous linear-by-linear association model A_L of Section 4 can be written as $\log\pi_{kij} = \tau_k + \tau_{ki}^{(1)} + \tau_{kj}^{(2)} + \lambda_k \cdot i \cdot j$, where, defining $\boldsymbol{\omega}_k \equiv (\tau_{k1}^{(1)}, \tau_{k2}^{(1)}, \ldots, \tau_{k4}^{(2)}, \lambda_k)^T$ and $\mathbf{A}_k = [0, 0, \ldots, 0, 1]$, $\lambda_k = \mathbf{A}_k\boldsymbol{\omega}_k = g_k(\alpha) \equiv \alpha, \quad k = 1, 2.$

The link $\mathbf{L}_k^*$ is one-to-one, but its domain lies in a lower dimension than its range. This "problem" can be avoided by reparameterizing the multinomial log likelihood $\sum_k \sum_{ij} n_{kij} \log \pi_{kij}$ in terms of $\boldsymbol{\omega}_k$ through the one-to-one transformation $\boldsymbol{\pi}_k = \boldsymbol{\pi}_k(\boldsymbol{\omega}_k) \equiv \exp(\mathbf{W}_k\boldsymbol{\omega}_k)/(\mathbf{1}^T \exp(\mathbf{W}_k\boldsymbol{\omega}_k))$. Using this reparameterized log likelihood, it can be shown that the BAM model (6.1) can be equivalently expressed as

$$\mathbf{L}_k(\boldsymbol{\omega}_k) \equiv \begin{bmatrix} \mathbf{M}_k\boldsymbol{\pi}_k(\boldsymbol{\omega}_k) \\ \mathbf{A}_k\boldsymbol{\omega}_k \end{bmatrix} = \begin{bmatrix} \mathbf{f}_k(\boldsymbol{\beta}) \\ \mathbf{g}_k(\boldsymbol{\alpha}) \end{bmatrix} \equiv \mathbf{h}_k(\boldsymbol{\theta}), \tag{6.2}$$

where the link $\mathbf{L}_k$ is one-to-one with non-singular derivative matrix. It is of practical interest to note that for parsimonious association models, reparametrizing the multinomial log likelihood in terms of $\boldsymbol{\omega}_k$ can mean a significant reduction in the size of the fitting problem.

Because $\mathbf{L}_k$ is invertible, the multinomial log likelihood $\sum_k \sum_{ij} n_{kij} \log \pi_{kij}(\boldsymbol{\omega}_k)$ can again be reparameterized, this time in terms of $\boldsymbol{\theta}$. This is accomplished by replacing $\boldsymbol{\omega}_k$ with $\mathbf{L}_k^{-1}(\mathbf{h}_k(\boldsymbol{\theta}))$. If $\mathbf{L}_k$ could be analytically inverted, we could use a standard Fisher-scoring algorithm to find the maximum likelihood estimate of $\boldsymbol{\theta}$ and an approximate variance estimate. Unfortunately, the link $\mathbf{L}_k$ generally cannot be analytically inverted. It must be numerically inverted, a task that can be accomplished using a Newton-Raphson algorithm. It follows that a two-stage nested algorithm similar to that of Glonek and McCullagh (1995) or Glonek (1996) can be used. The "outside" iterations update $\boldsymbol{\theta}$ estimates, while the "inside" iterations are used to invert the link. Although, Glonek and McCullagh (1995) and Glonek (1996) considered different link functions and $\mathbf{h}_k$ functions that were linear in $\boldsymbol{\theta}$, their algorithms can be modified for applicability in this more general non-linear case. The appendix gives a more formal outline of the maximum likelihood iterative fitting algorithm used in BAMROC.

This section closes with a discussion of ML estimation of AUC's and their differences. By standard ML theory, $\hat{\boldsymbol{\theta}}$ has an approximate normal distribution with variance matrix $avar(\hat{\boldsymbol{\theta}}) = \mathbf{I}^{-1}(\hat{\boldsymbol{\theta}})$, the inverse of the estimated expected information matrix. One important benefit of using the Fisher-scoring fitting algorithm as outlined in the appendix is that $avar(\hat{\theta})$ is obtained as a side effect.

For binormal models, the area under the m^{th} (latent) ROC curve is

$$AUC_m = \Phi(\frac{\mu_m}{\sqrt{1+\mathrm{e}^{2\xi_m}}}),$$

where Φ is the standard normal distribution function. By invariance properties, the maximum likelihood estimator is $A\hat{U}C_m = \Phi(\frac{\hat{\mu}_m}{\sqrt{1+\mathrm{e}^{2\hat{\xi}_m}}})$, which is a differentiable function of $\hat{\theta}$. Similarly, the difference $A\hat{U}C_1 - A\hat{U}C_2$ is a differentiable function of $\hat{\theta}$. Because $avar(\hat{\theta})$ is computed as part of the fitting algorithm, approximate (asymptotic) standard errors of the $A\hat{U}C$ estimators and their difference can be easily computed using the delta method.

7 Analysis of Neonatal Radiograph Data

Let A_I and A_L represent the independence and homogeneous linear-by-linear association model, respectively, as specified in Section 4.

We fitted three candidate models to the data of Section 2, namely the bivariate binormal (BB), BA_IM, and BA_LM. The BB model was fitted using the program CORROC2 (see Metz et al. 1984) available at Dr. Charles E. Metz's website. Select results are given in Table 2. The symbol G^2 represents the likelihood-ratio-statistic for testing goodness of fit of the model; $ase(DIFF)$ is the approximate standard error for the difference between the two area estimators; and $\hat{\rho}_1$ and $\hat{\rho}_2$ are the latent suspicion score correlations for normals and abnormals, respectively.

Table 2. Maximum Likelihood Fit Results for Neonatal Radiograph Data

Model	G^2	df	$A\hat{U}C_1$	$A\hat{U}C_2$	$ase(DIFF)$	$\hat{\lambda}$ (ase)	$\hat{\rho}_1$ (ase)	$\hat{\rho}_2$ (ase)
BA_IM	52.51	36	0.863	0.861	0.052	–	–	–
BA_LM	29.46	35	0.862	0.858	0.038	0.386 (0.093)	–	–
BB	25.43	34	0.864	0.861	0.038	–	0.513 (0.151)	0.616 (0.104)

On the one hand, point estimation of the ROC areas is not very sensitive to choice of association structure. Standard error estimation, on the other hand, is sensitive to choice of association. For these data, both BA_LM and BB give reasonable overall fits, and inferences about the difference between areas under the ROC curves would be similar. In contrast, the poor-fitting independence association model BA_IM, which ignores correlation between the two ratings, leads to an inflated estimate of the standard error of the difference between areas.

Because of zero counts, the saturated association model could not be fit (and hence $G^2(BM)$, the likelihood-ratio-statistic for testing goodness of fit of the binormal marginal model, could not be calculated) without adding small constants. Alternative to adding constants, we conjecture that, because A_L fits reasonably well ($G^2(A_L) = 27.63, df = 31$), $G^2(BM) \approx G^2(BA_LM) - G^2(A_L) = 1.83, df = 4$ (cf. Lang et al. 1999). Apparently, the binormal assumption is tenable. Recall that the binormal assumption cannot be tested using the BB model.

Using the BA_LM model, we find that the estimated areas under the ROC curves are not statistically different. That is, the diagnostic performances of the video and plain film images, as measured using the "area index," are not statistically different. To detect other possible differences in diagnostic performance, we present in Figure 1 a graph of the fitted (under BA_LM) ROC curves and the empirical ROC points. That the ROC points all lie above the 45° line implies that empirically Y_m is stochastically larger than X_m; i.e. both modalities have diagnostic capabilities. That the ROC curves cross, might lead us to question the use of the area index as the only measure of diagnostic performance (cf. Metz et al. 1984).

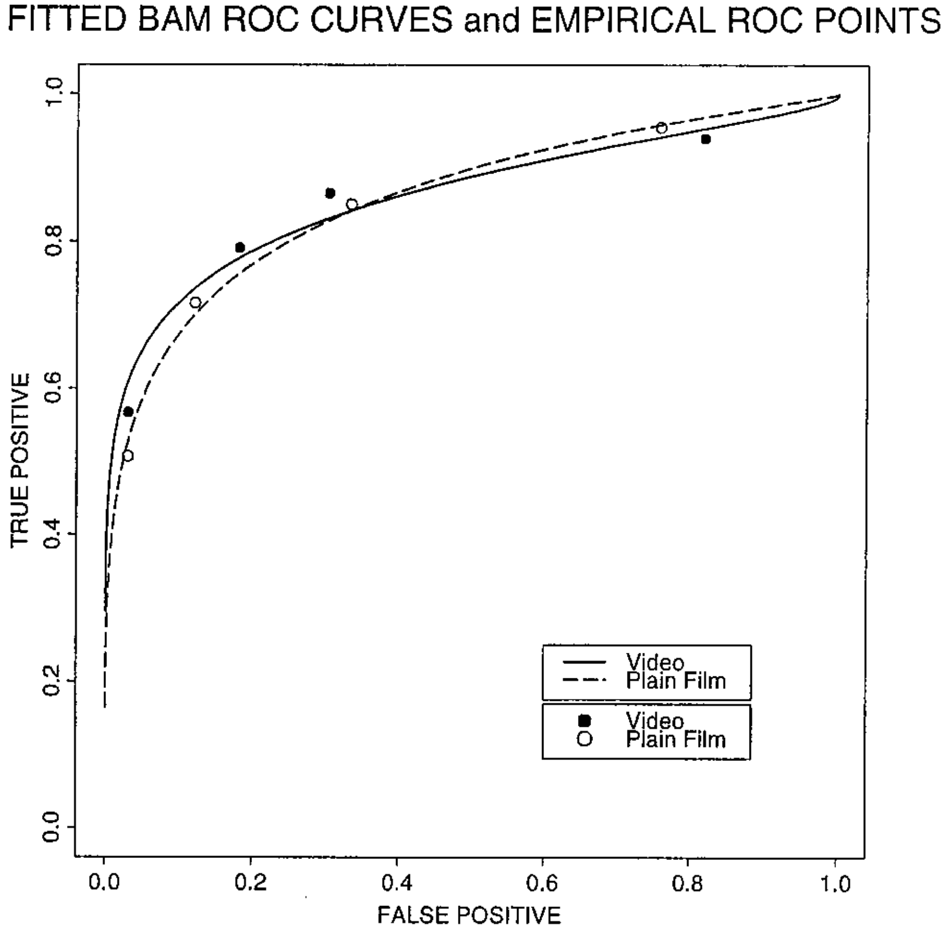

Figure 1. Fitted BA_LM ROC Curves and Empirical ROC Points.

The BA_LM model, which has association model of the form $A_L : \log \pi_{kij} = \tau_k + \tau_{ki}^{(1)} + \tau_{kj}^{(2)} + \lambda \cdot i \cdot j$, affords a very simple description of the association between the manifest suspicion scores for the two modalities. The model implies that the strength of the association as measured by the single parameter λ is the same across the control and case populations. This along with the choice of equal-interval scores implies a uniform association (cf. Agresti 1990) in that all of the local odds ratios, 16 from the control population and 16 from the case population, are identical. The ML estimate of the common local odds ratio value is $\exp(\hat{\lambda}) = \exp(0.386) = 1.47$ (ase = 0.093). In words, for either population, the odds of assigning a suspicion score of $i + 1$ rather than

i using Plain Film is estimated to be 1.47 [95% confidence interval: 1.22, 1.77] times higher when the Video suspicion score is $j + 1$ than when it is j. As expected there is a significant positive association.

8 A Small Simulation Study

This section illustrates the broad applicability of the class of BAM models.

The following tables (see Table 3) were generated from a model with binormal margins (a common AUC of 0.78 was used) and quasi-independence association structure (see Model A_Q of Section 4). Thus, a BAM model holds, but the bivariate binormal model does not.

Table 3. Simulated BAM data.

		Mod II					
		1	2	3	4	5	
Mod I	1	14	0	3	0	1	18
	2	0	3	0	1	0	4
	3	2	1	2	0	0	5
	4	1	0	0	0	0	1
	5	2	0	2	0	1	5
		19	4	7	1	2	33

Controls

		Mod II					
		1	2	3	4	5	
Mod I	1	8	0	1	0	1	10
	2	0	3	0	0	0	3
	3	0	0	10	0	1	11
	4	0	2	1	9	1	13
	5	0	2	1	1	26	30
		8	7	13	10	29	67

Cases

We fitted the bivariate binormal (BB), BA_IM, and BA_QM to these generated data. Select results are given in Table 4.

Table 4. Maximum Likelihood Fit Results for BAM Simulated Data.

Model	G^2	df	$A\hat{U}C_1$ (ase)	$A\hat{U}C_2$ (ase)	$ase(DIFF)$
BA_IM	152.15	36	0.787 (0.053)	0.835 (0.044)	0.069
BA_QM	30.38	34	0.783 (0.053)	0.835 (0.044)	0.051
BB	85.00	34	0.786 (0.053)	0.828 (0.044)	0.050

As before, notice that point estimates of areas are not sensitive to choice of association structure. The standard error estimate of the difference between area estimators for the BA_IM model, which fits poorly, is saliently inflated relative to the standard errors for the other two models, which allow for correlation between responses. For these data, the BB and BA_IM models fit poorly overall, while the BA_QM model fits well (as it should). While the association between the two ordinal responses is easy to describe using the good-fitting BA_QM model, a reasonable description is not available using the poor-fitting BB model. While the binormal assumption cannot be tested using BB, it can be with BA_QM. In fact, $G^2(BM) \approx G^2(BA_QM) - G^2(A_Q) = 30.38 - 25.58 = 4.80$ ($df = 4$), and the binormal assumption is (correctly) deemed tenable.

This simulated example illustrates that there are BAM data for which the bivariate binormal model is too restrictive and does not fit well. In contrast, bivariate binormal data can always be well-fitted using a BAM model with unrestricted association structure, assuming there are no problems with zero counts. In fact, by arguments of Goodman (1981) and Becker (1989), there are *parsimonious* BAM models, for example BAM models with linear-by-linear association structures, that will typically fit bivariate binormal data well.

In sum, a good-fitting model is especially important if the scientific objective includes describing the association structure. The bivariate binormal models can be overly restrictive and may fit data poorly. In contrast, most data can be well-modeled using members of the rich class of BAM models. As a special case, this rich class includes models that fit bivariate binormal data very well.

9 Discussion

BAM models impose the well-understood and commonly-accepted binormal structure of classic signal detection theory to each ordinal response marginal distribution. Therefore, unlike non-parametric models, BAM models afford measures of diagnostic performance that are independent of the way that the ordinal rating scales are defined (see Section 3). In contrast to bivariate binormal models, BAM models do not impose the added restriction that the joint distributions are determined by latent bivariate normal distributions. Instead of explicitly specifying the joint distributions of the latent responses, BAM models directly specify the ordinal variable association structure using standard loglinear models. It follows that BAM models are very flexible and easy to interpret. They are particularly useful when both the ordinal response marginal and association structures are of scientific interest.

One should exercise caution when evaluating the overall goodness of fit of a BAM model to sparse tables. In particular, large-sample chi-squared approximations to the null distribution of the overall goodness-of-fit statistics can be very poor. In the sparse table setting, it is recommended that a parsimonious, theoretically-viable association model be tested against a more general, but parsimonious, association model. Failure to reject the simpler model should lend credence to its use. Alternatively, Markov chain Monte Carlo exact conditional methods for testing goodness of fit of a loglinear association model could be employed (Forster et al. 1996)

Upon more careful inspection of the ML fitting algorithm used in BAMROC as outlined in the Appendix, one can see that it is applicable for a much broader class of models than the BAM models considered herein. For example, the marginal model functions $\mathbf{f}_k$ can be quite general; they needn't be restricted to functions of normal probabilities. For example, bi-gamma models (Dorfman et al. 1997) for the marginals could be used instead of binormal models. The association structure need not be loglinear, because the association model functions $\mathbf{g}_k$'s can be non-linear in their arguments. As an example, log-bilinear RC association models (cf. Goodman 1985) could be used.

An S-plus version of the BAMROC program used to analyze the examples in this paper, and supporting documentation, can be obtained from the second author (`aspelund@stat.uiowa.edu`).

Appendix: The Iterative Scheme used in BAMROC

The goal is to maximize the log likelihood

$$\ell(\boldsymbol{\theta}) \equiv \sum_k \mathbf{n}_k^T \log \boldsymbol{\pi}_k(\boldsymbol{\omega}_k(\boldsymbol{\theta})) \equiv \sum_k \ell_k(\boldsymbol{\omega}_k(\boldsymbol{\theta})),$$

with respect to $\boldsymbol{\theta}$, where $\boldsymbol{\omega}_k(\boldsymbol{\theta})$ is implicitly defined through $\mathbf{L}_k(\boldsymbol{\omega}_k) = \mathbf{h}_k(\boldsymbol{\theta})$, and $\mathbf{L}_k$ and $\mathbf{h}_k$ are as defined in Sections 4 and 2.

Letting $\mathbf{s}(\boldsymbol{\theta}) = \partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ be the score, the maximizer will generally solve the likelihood equations $\mathbf{s}(\boldsymbol{\theta}) = \mathbf{0}$. The Fisher-scoring algorithm is used to solve these equations. Specifically, for a current iterate $\boldsymbol{\theta}$, the updating equation has the form:

$$\boldsymbol{\theta}_{new} = \boldsymbol{\theta} + \mathbf{I}(\boldsymbol{\theta})^{-1}\mathbf{s}(\boldsymbol{\theta}), \tag{A.1}$$

where $\mathbf{I}(\boldsymbol{\theta})$ is the Fisher expected information matrix.

The score vector $\mathbf{s}$ has the form

$$\mathbf{s}(\boldsymbol{\theta}) = \sum_k \frac{\partial\ell_k(\boldsymbol{\omega}_k(\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}} = \sum_k \left(\frac{\partial\boldsymbol{\omega}_k(\boldsymbol{\theta})^T}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial\ell_k(\boldsymbol{\omega}_k)}{\partial\boldsymbol{\omega}_k}|_{\boldsymbol{\omega}_k=\boldsymbol{\omega}_k(\boldsymbol{\theta})}\right),$$

and the expected Fisher information can be shown to have the form

$$\mathbf{I}(\boldsymbol{\theta}) \quad = -E(\frac{\partial\mathbf{s}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T}) = \sum_k n_{k++}\left(\frac{\partial\boldsymbol{\omega}_k(\boldsymbol{\theta})^T}{\partial\boldsymbol{\theta}}\right)\mathbf{W}_k^T[\mathbf{D}(\boldsymbol{\pi}_k) - \boldsymbol{\pi}_k\boldsymbol{\pi}_k^T]\mathbf{W}_k\left(\frac{\partial\boldsymbol{\omega}_k(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T}\right),$$

where $\mathbf{D}(\boldsymbol{\pi}_k)$ is the diagonal matrix with components in $\boldsymbol{\pi}_k$ on the diagonal and $\boldsymbol{\pi}_k = \boldsymbol{\pi}_k(\boldsymbol{\omega}_k(\boldsymbol{\theta}))$. Also, using the fact that $\mathbf{L}$ has a non-singular derivative matrix, we have that

$$\frac{\partial\boldsymbol{\omega}_k(\boldsymbol{\theta})^T}{\partial\boldsymbol{\theta}} = \left(\frac{\partial\mathbf{h}_k(\boldsymbol{\theta})^T}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial\mathbf{L}_k(\boldsymbol{\omega}_k)^T}{\partial\boldsymbol{\omega}_k}|_{\boldsymbol{\omega}_k=\boldsymbol{\omega}_k(\boldsymbol{\theta})}\right)^{-1}.$$

For the current value of $\boldsymbol{\theta}$, *the value of the implicitly-defined function* $\boldsymbol{\omega}_k(\boldsymbol{\theta})$ *is determined.* Given both $\boldsymbol{\theta}$ and $\boldsymbol{\omega}_k(\boldsymbol{\theta})$, all the derivatives involved in the calculation of $\mathbf{s}(\cdot)$ and $\mathbf{I}(\cdot)$ can be computed; the $\mathbf{h}_k$ derivatives are computed numerically and the others are computed analytically. Given the values of these derivatives, the Fisher-scoring update (A.1) is readily obtained.

The italicized phrase in the previous paragraph indicates that the Fisher-scoring update requires computation of $\boldsymbol{\omega}_k(\boldsymbol{\theta})$ for the given value of $\boldsymbol{\theta}$. But $\boldsymbol{\omega}_k(\boldsymbol{\theta})$ equals the value of $\boldsymbol{\omega}_k$ that solves the equation $\mathrm{L}_k(\boldsymbol{\omega}_k) = \mathrm{h}_k(\boldsymbol{\theta})$, and this solution generally cannot be obtained analytically. Owing to the form of the model, this root-finding problem, i.e. link inversion, can be simplified somewhat. Specifically, the loglinear parameter $\boldsymbol{\omega}_k$ can be partitioned as $\boldsymbol{\omega}_k = (\boldsymbol{\eta}_k, \boldsymbol{\lambda}_k)$, where $\boldsymbol{\lambda}_k \equiv \mathrm{A}_k\boldsymbol{\omega}_k$ comprises the association parameters and $\boldsymbol{\eta}_k$ comprises the main-effects parameters. Now, the $\boldsymbol{\omega}_k$ solution satisfies $\boldsymbol{\lambda}_k = \mathrm{g}_k(\alpha)$, which is known because the current iterate value $\theta = (\alpha, \beta)$ is given. Therefore, the $\boldsymbol{\omega}_k$ solution can be obtained, i.e. the link inverted, once we solve for $\boldsymbol{\eta}_k$ in the reduced set of equations $\mathrm{M}_k\boldsymbol{\pi}^*_k(\boldsymbol{\eta}_k) = \mathrm{f}_k(\beta)$, where $\boldsymbol{\pi}^*_k(\boldsymbol{\eta}_k) = \boldsymbol{\pi}_k(\boldsymbol{\eta}_k, \mathrm{g}_k(\alpha))$. We use a Newton-Raphson iterative scheme to solve this reduced system of equations. It is important to note that this simplification can represent a significant reduction in the size of the link-inversion problem. This implies that the link inversion is faster and problems with zero counts are mitigated. This link-inversion approach is related to that of Glonek (1996).

References

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

Bamber, D. (1975) The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, **12**, 387-415.

Becker, M.P. (1989) On the bivariate normal distribution and association models for ordinal categorical data. *Statistics & Probability Letters*, 435-440.

Delong, E.R., Delong, D.M. and Clarke-Pearson, D.L. (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44, 837-845.

Dorfman, D.D. and Alf, Jr., E. (1969) Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—rating method data. *Journal of Mathematical Psychology*, **6**, 487-496.

Dorfman, D.D., Berbaum, K.S., and Metz, C. E. (1992) Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. *Investigative Radiology*, **27**, 723-731.

Dorfman D.D., Berbaum K.S., Metz C.E., Lenth R.V., Hanley J.A., AbuDagga H. (1997) Proper receiver operating characteristic analysis: The bigamma model. *Academic Radiology,* **4**, 138-149.

Forster, J.J., McDonald, J.W., and Smith, P.W.F. (1996) Monte Carlo exact conditional tests for log-linear and logistic models. *Journal of the Royal Statistical Society,* Ser. B, **58**, 445-453.

Glonek, G.F.V. (1996) A class of regression models for multivariate categorical responses. *Biometrika*, **83**, 15-28.

Glonek, G.F.V. and McCullagh, P. (1995) Multivariate logistic models. *Journal of the Royal Statistical Society-Series B*, **57**, 533-546.

Goodman, L.A. (1981) Association models and the bivariate normal for contingency tables with ordered categories. *Biometrika*, **68**, 347-55.

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.

Green, D.M. and Swets, J.A. (1966) *Signal detection theory and psychophysics.* New York: Wiley.

Hanley, J.A. (1998) Receiver operating characteristic (ROC) curves. In Armitage, P. and Colton, T., eds. *Encyclopedia of Biostatistics, Vol. 4.* New York: Wiley, 3738-3745.

Hanley, J. A. and McNeil, B. J. (1982) The meaning and use of the area under a Receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29-36.

Hanley, J. A. and McNeil, B. J. (1983) Method for comparing the area under the ROC curves derived from the same cases. *Radiology*, **148**, 839-843.

Ishwaran, H. and Gatsonis, C.A. (2000) A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. Forthcoming in *The Canadian Journal of Statistics*, **28**, No. 4.

Metz, C. E. (1978) Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, **8**, 283-298.

Metz, C. E. and Kronman, H. B. (1980) Statistical significance tests for binormal ROC curves. *Journal of Mathematical Psychology*, **22**, 218-243.

Metz, C. E. and Wang, P. L. and Kronman, H. B. (1984) A new approach for testing the significance of differences measured from correlated data. In Deconinck, F., ed. *Information Processing in Medical Imaging VIII.* Dordrecht, the Netherlands: Martinus Nijhoff, 432-445.

Obuchowski, N. A. (1997) Nonparametric analysis of clustered ROC curve data. *Biometrics*, **53**, 567-578.

Swets, J. A. and Pickett, R. M. (1982) *Evaluation of diagnostic systems: methods from signal-detection theory.* New York: Academic Press.

Thompson, M. L. and Zucchini, W. (1989) On the statistical analysis of ROC curves. *Statistics in Medicine*, **8**, 1277-1290.

Toledano, A. and Gatsonis, C. A. (1995) Regression analysis of correlated receiver operating characteristic data. *Academic Radiology*, **2**, S30-S36.

Toledano, A. and Gatsonis, C. A. (1996) Ordinal regression methodology for ROC curves derived from correlated data. *Statistics in Medicine*, **15**, 1807-1826.

ACKNOWLEDGMENTS: *The authors would like to thank Dr. Donald Dorfman for many helpful discussions and comments.*