The Canadian Journal of Statistics
Vol. 28, No. 4, 2000, Pages 731–750
La revue canadienne de statistique

A general class of hierarchical ordinal regression models with applications to correlated ROC analysis

Hemant ISHWARAN and Constantine A. GATSONIS

Key words and phrases: Bayesian hierarchical model; Gibbs sampling; HROC model; ordinal regression; ordinal categorical data; ROC curve.

AMS 1991 subject classifications: Primary 62H99; secondary 62F99, 62P10.

ABSTRACT

The authors discuss a general class of hierarchical ordinal regression models that includes both location and scale parameters, allows link functions to be selected adaptively as finite mixtures of normal cumulative distribution functions, and incorporates flexible correlation structures for the latent scale variables. Exploiting the well-known correspondence between ordinal regression models and parametric ROC (Receiver Operating Characteristic) curves makes it possible to use a hierarchical ROC (HROC) analysis to study multilevel clustered data in diagnostic imaging studies. The authors present a Bayesian approach to model fitting using Markov chain Monte Carlo methods and discuss HROC applications to the analysis of data from two diagnostic radiology studies involving multiple interpreters.

RÉSUMÉ

Les auteurs s'intéressent à une classe assez vaste de modèles de régression ordinale avec paramètres de localisation et d'échelle, laquelle permet la sélection adaptative de fonctions de lien s'exprimant comme mélanges finis de fonctions de répartition normales et fournit des structures de corrélation flexibles pour les variables d'échelle latentes. En exploitant la correspondance bien connue entre les modèles de régression ordinale et les courbes d'efficacité paramétriques (CEP) des tests diagnostiques, il est possible d'analyser des données d'imagerie médicale diagnostique regroupées à plusieurs niveaux au moyen d'une CEP hiérarchique. Les auteurs décrivent une approche bayésienne pour l'ajustement de tels modèles au moyen des méthodes de Monte Carlo à chaîne de Markov et présentent deux applications concrètes concernant l'interprétation de clichés radiologiques.

1. INTRODUCTION

Multilevel ordinal categorical data structures arise in research settings in which clustered and/or longitudinal ordinal categorical responses are observed. In diagnostic radiology studies, ordinal categorical response data are collected from the radiologists' degree of suspicion of abnormality on a patient's radiology scan. In practice, the degree of suspicion is usually recorded on a 5-point ordinal scale, which is typically encoded in a set of verbal descriptors such as: "definitely normal," "probably normal," "equivocal," "probably abnormal," and "definitely abnormal." However, we note that more refined scales are also used in practice, with some researchers proposing the use of ordinal scales with as many as 100 categories (Rockette, Gur & Metz 1992). The broad range of experimental designs used in the diagnostic study leads to a number of ways in which the ordinal data become clustered. For example, some studies require each patient to be imaged using different diagnostic modalities, with the images then being interpreted independently by several different radiologists from different participating institutions. This type of design leads to correlated ordinal data that are clustered at the level of the institution. In other studies, including in meta-analysis of diagnostic test evaluations, patients are nested completely within radiologists, or they are nested within the study. In addition to the ordinal categorical responses, the data in diagnostic studies often include patient- and cluster-level covariates of interest, such as true dis-

ease status and other clinical information for patients, experience and training for radiologists, and institutional characteristics for hospitals. The goal of the analysis of this type of data is to assess diagnostic accuracy and to examine its relation to patient, reader, and possibly institutional characteristics.

A commonly used method for assessing the accuracy of a diagnostic test is through what is called an ROC (Receiver Operating Characteristic) analysis; a method developed to account for the inherent trade-off between the sensitivity and specificity of a diagnostic test. The theoretical ROC curve of a diagnostic test is the plot of all possible pairs of values for the test's false positive rate (1-specificity) and true positive rate (sensitivity), while the empirical ROC curve plots the observed pairs as the threshold for the test is varied over the degree of suspicion categories. A smooth ROC curve can be obtained by fitting a parametric model to the ordinal categorical data and letting the threshold for test positivity vary over the entire range of the latent "degree of suspicion" variable. As discussed in the next section, the ordinal regression model with a cumulative link (McCullagh 1979, 1980) can be used to perform regression analysis of ROC curves.

The focus of this paper is on the use of Bayesian hierarchical ordinal regression models in ROC analysis. We shall use the abbreviation HROC (Hierarchical ROC) for this type of analysis. The HROC models developed in this paper can include covariates reflecting characteristics of the units at various levels of aggregation, such as individual patients, radiologists, and hospitals. These models make it possible to develop ROC curves for groups of patients and individual radiologists, to define "average" curves for diagnostic modalities, and to study the variability across radiologists interpreting a diagnostic modality. In addition, the models make it possible to estimate a correlation structure for the ordinal response as a method for studying the consistency of the response over multiple image interpretations.

The HROC models discussed in this paper build on the cumulative link regression model pioneered by McCullagh (1979, 1980). In its simplest form, the cumulative link model relates the observed ordinal response to covariates through an unobserved continuous latent construct and a prespecified monotone link function. The model may contain both location and scale parameters, and was originally used to carry out fixed effects analysis of data in which each subject has only one observation. Harville & Mee (1984) later extended this model to include a random effects formulation. Later, the problem of multiple response data was studied by Stram, Wei & Ware (1988), Uebersax (1993), Uebersax & Grove (1993) and Hedeker & Gibbons (1994). A generalized estimating equations approach to the problem has also been discussed in Toledano (1993), Williamson, Kim & Lipsitz (1995) and Toledano & Gatsonis (1995).

Bayesian methodology based on the unobserved latent variable approach was pioneered by Albert & Chib (1993) for the single ordinal response model, with extensions given later by Erkanli, Stangl & Müller (1993). For the multiple response problem, Gatsonis (1995) studied the use of random effects in radiology data, Johnson (1996) presented a multiple rater application to essay grading, and Cowles, Carlin & Connett (1996) presented a multivariate Tobit analysis for the problem of nonignorable missing data. Our contribution to this literature is the description of a flexible class of models appropriate for HROC analysis, developed by integrating various technical features of the Bayesian ordinal regression model. In particular, some key components of the models to be discussed are that they allow for a range of patterns of clustering with covariates at each level of the hierarchical structure, they permit adaptive link selection, and they allow for more flexible modeling of the correlation in the data. Although simpler versions of these features have been considered separately in some of the methodologic literature on ordinal and binary data (Chib & Greenberg 1998), we use a combination of these features at a higher level of technical complexity, which is made necessary by the analytic needs of diagnostic imaging data. Because of the level of complexity, the resulting HROC analysis discussed in this paper is considerably more general than other published work in this area (Gatsonis 1995, Hellmich, Abrams, Jones & Lambert 1998).

Section 2 gives a brief description of the traditional ordinal regression model, while the discussion of the HROC model is given in Section 3. The details of model fitting via our Gibbs sampler are presented in the Appendix. Section 4 contains an application to data from a prostate cancer study involving a 100-point scale and several radiologists. In Section 5, we study data from a recent radiology study which utilized a conventional 5-point scale and involved over 20 radiologists, several hospitals and two different imaging methods. The two examples illustrate two types of scales typically seen: the usual 5-point scale and a 100-point scale. They also illustrate the different types of clustering that naturally arise in multi-reader radiology studies: clustering by radiologist (Section 4) and clustering by institution and type of diagnostic modality (Section 5). Finally, in Section 6 we present simulations to test the sensitivity of the proposed models. Section 7 contains a discussion.

2. ORDINAL REGRESSION MODEL

The ordinal regression model proposed by McCullagh (1979, 1980) assumes that independent ordinal categorical observations Y_i and location covariates $\mathbf{x}_i$ and scale covariates $\mathbf{u}_i$ are available on cases $i = 1, \ldots, N$. The response $Y_i \in \{r_1, \ldots, r_J\}$ is assumed to result from the classification of some unobserved continuous latent variable, with the implicit ordering in the response outcomes (r_1 is the "smallest" categorical value and r_J is the "largest" categorical value) being related to unknown cutpoints $\theta_1, \ldots, \theta_{J-1}$ which form a set of contiguous intervals for the underlying latent scale. The cumulative link ordinal regression model specifies that

$$\mathrm{P}(Y_i \le r_j \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{x}_i, \mathbf{u}_i) = h\left\{\frac{\theta_j - \boldsymbol{\beta}'\mathbf{x}_i}{\exp(\boldsymbol{\alpha}'\mathbf{u}_i)}\right\}, \qquad \text{for } j = 1, \ldots, J, \tag{1}$$

where $h\colon \mathbb{R} \to (0,1)$ is a differentiable monotone link function, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are unknown vectors of scale and location parameters and $\boldsymbol{\theta}$ is the vector of cutpoints whose coordinates satisfy

$$-\infty = \theta_0 \le \theta_1 \le \cdots \le \theta_{J-1} \le \theta_J = +\infty. \tag{2}$$

The model has a particularly intuitive interpretation when the link function is a standard normal cdf. In this case, the ordinal responses are classified according to

$$Y_i = r_j \qquad \text{if and only if} \qquad \theta_{j-1} < M_i \le \theta_j,$$

where

$$M_i = \boldsymbol{\beta}'\mathbf{x}_i + Z_i \exp(\boldsymbol{\alpha}'\mathbf{u}_i)$$

is a latent (unobserved) normal variable and $Z_i \sim \mathrm{N}(0,1)$. In other words, the ordinal response variables are assumed to be derived as a classification of a normal variable with mean $\boldsymbol{\beta}'\mathbf{x}_i$ and variance $\exp(2\boldsymbol{\alpha}'\mathbf{u}_i)$. In principle, the distribution for Z_i could be replaced by any normal distribution as long as its mean and variance are fixed in advance to ensure identification for parameters. The use of a standard normal distribution is the preferred choice for convenience, and leads to a clear interpretation for the parameters.

The ordinal regression model provides an effective approach for a regression analysis of parametric ROC curves. As noted in the discussion of McCullagh (1980), and later elucidated by Tosteson & Begg (1988), the binormal ROC model (that is, the ROC model in which the latent variable is assumed to be normally distributed) is equivalent to an ordinal regression model with a probit link and a single binary covariate $x_i = u_i$ indicating true disease status ($-1/2$ = disease, $1/2$ = normal). Then, β_1 measures the diagnostic procedure's ability to discriminate between the diseased and nondiseased groups, while the scale parameter α_1 describes the change in the variability of the latent variable across normal and diseased patients. In particular, if α_1 is fixed, a large positive value for β_1 indicates a more accurate test, while a positive value for α_1 indicates a test with less variability for the non-diseased cases than for diseased cases.

An often used summary value measuring the accuracy of a diagnostic test is the area under the ROC curve (see Hanley & McNeil 1982). In this approach, the effect of a covariate on diagnostic performance is determined by the resulting ROC curve and its area. Thus, for example, to compute the area under the empirical ROC curve, and hence determine the accuracy for a specific covariate $(\mathbf{x}, \mathbf{u})$, we construct the ROC curve for the covariate by plotting the test's false positive rate against its true positive rate as the response r_j changes. Whatever the choice for the link function, this ROC curve is derived by plotting

$$\left(\mathrm{P}\{Y \geq r_j \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{x}^{(1)}, x_1 = -1/2, \mathbf{u}\}, \mathrm{P}\{Y \geq r_j \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{x}^{(1)}, x_1 = 1/2, \mathbf{u}\}\right),$$

for $j = 1, \ldots, J$, where we have assumed for simplicity that $\mathbf{u}$ contains no information about disease and that $\mathbf{x}$ is defined so that its first coordinate x_1 records true disease status and $\mathbf{x}^{(1)}$ denotes the sub-vector of $\mathbf{x}$ with x_1 removed.

The smoothed ROC curve is derived in a similar fashion as in the previous display, but with the categorical values r_j replaced by a variable which is then varied over the entire range of the latent variable. An alternative formulation of ROC curves, which also leads to regression analysis for independent and correlated data was discussed in a recent paper by Pepe (1997). A discussion of linear model methods for ROC curve summaries, such as the area under the curve, can be found in Thompson & Zucchini (1989) and Obuchowski (1995).

3. HIERARCHICAL ROC MODELS

The ordinal regression model (1) can be used as the basic building block for hierarchical models suitable for the analysis of multilevel clustered ROC data. In this setting, we observe k_i ordinal categorical responses $Y_{i,1}, \ldots, Y_{i,k_i}$ for each individual $i = 1, \ldots, N$, where the $Y_{i,k}$ represent repeated observations for the subject i collected from a maximum of K different clusters ($1 \leq k_i \leq K$). In addition to the ordinal response data, there is also covariate information available, which in some cases can be patient specific only (depending only upon i), and in other cases it is both patient and cluster specific (depending upon i and k).

Following the style of the univariate model (1), we consider covariates as either location or scale specific. Therefore, we have covariates $\mathbf{x}_i$ and $\mathbf{x}_{i,k}$ for location parameters β_0 and $\boldsymbol{\beta}$, and we have covariates $\mathbf{u}_i$ and $\mathbf{u}_{i,k}$ for scale parameters α_0 and $\boldsymbol{\alpha}$. In the analysis of diagnostic studies, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are typically made up of a collection of cluster specific parameters, such as parameters for different readers, or for different hospitals. These sets of cluster specific parameters are then identified by using the covariates $\mathbf{u}_{i,k}$ and $\mathbf{x}_{i,k}$ as vectors containing components that function as indicator terms. In contrast, the covariates $\mathbf{u}_i$ and $\mathbf{x}_i$ are fixed for each patient i, and represent patient clinical covariates for parameters α_0 and β_0.

Our model is based on the following latent construct description:

$$Y_{i,k} = r_j \qquad \text{if and only if} \qquad \theta_{j-1} < M_{i,k} \leq \theta_j, \tag{3}$$

for $j = 1, \ldots, J$, where θ_j are cutpoints (2) and

$$M_{i,k} = \beta_0' \mathbf{x}_i + \boldsymbol{\beta}' \mathbf{x}_{i,k} + (Z_{i,k} + \delta_{i,k}) \exp(\alpha_0' \mathbf{u}_i + \boldsymbol{\alpha}' \mathbf{u}_{i,k}) \tag{4}$$

is a latent variable with a joint distribution depending upon $\mathbf{Z}_i = (Z_{i,1} \ldots, Z_{i,k_i})'$. We will assume that

$$\mathbf{Z}_i \mid \mathbf{R} \sim \mathrm{N}(\mathbf{0}, \mathbf{R}_i), \tag{5}$$

where $\mathbf{R}$ is the correlation matrix for the K clusters (dimension $K \times K$) and $\mathbf{R}_i$ is the $k_i \times k_i$ sub-correlation matrix of the responses for subject i. Note that we work with a correlation matrix in specifying the distribution for $\mathbf{Z}_i$, because just as in the univariate setting of Section 2, the model is unidentified without fixing the variances for $Z_{i,k}$ in advance. As mentioned, a particularly convenient choice in the univariate model is to set the variances equal to 1, which has the added

benefit in the multivariate model of forcing the covariance of $\mathbf{Z}_i$ to be a correlation matrix, providing a convenient interpretation for the correlation in the ordinal data.

Our approach also produces a generalized mixture link function, which is induced by the choice of priors for the discrete variables $\delta_{i,k}$ defined by

$$\mathrm{P}(\delta_{i,k} = \gamma_s \mid \mathbf{p}, \boldsymbol{\gamma}) = p_s, \qquad s = 1, \ldots, d \tag{6}$$

where

$$\mathbf{p} \mid \mathbf{a} \sim \mathrm{Dirichlet}_d(\mathbf{a}), \tag{7}$$

and $\boldsymbol{\gamma} = (\gamma_1 \ldots, \gamma_d)'$ is a prechosen grid of support points. As equations (8) and (9) below will show, this has the effect of producing a generalized link function for the conditional distribution of $Y_{i,k}$.

3.1. Hierarchical models.

The HROC models are derived by placing priors on the parameters α_0, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $\mathbf{R}$ in the latent construct formulation (3), (4) and (5). These class of models have the following general hierarchical description.

Level I (Joint distribution for the ordinal response). At the first stage of the hierarchy, we define the joint distribution for the ordinal response conditioned on parameters. We have (suppressing the dependence on covariates),

$$\begin{aligned}\mathrm{P}(Y_{i,1} = y_1, \cdots, Y_{i,k_i} = y_{k_i} \mid \alpha_0, \boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\delta}_i, \mathbf{R}) \\ = \int_{S_1} \int_{S_2} \cdots \int_{S_{k_i}} |2\pi \mathbf{R}_i|^{-1/2} \exp\Big\{-\frac{1}{2}(\mathbf{Z}_i - \boldsymbol{\delta}_i)' \mathbf{R}_i^{-1} (\mathbf{Z}_i - \boldsymbol{\delta}_i)\Big\}\, d\mathbf{Z}_i,\end{aligned}$$

where $\boldsymbol{\delta}_i = (\delta_{i,1}, \ldots, \delta_{i,k_i})'$ and

$$S_k = \left(\frac{\theta_{j-1} - \boldsymbol{\beta}_0' \mathbf{x}_i - \boldsymbol{\beta}' \mathbf{x}_{i,k}}{\exp(\boldsymbol{\alpha}_0' \mathbf{u}_i + \boldsymbol{\alpha}' \mathbf{u}_{i,k})}, \frac{\theta_j - \boldsymbol{\beta}_0' \mathbf{x}_i - \boldsymbol{\beta}' \mathbf{x}_{i,k}}{\exp(\boldsymbol{\alpha}_0' \mathbf{u}_i + \boldsymbol{\alpha}' \mathbf{u}_{i,k})}\right], \qquad \text{if } y_k = r_j.$$

In particular, this implies that $Y_{i,k}$ has the conditional distribution

$$\mathrm{P}(Y_{i,k} \le r_j \mid \alpha_0, \boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{p}) = h\left\{\frac{\theta_j - \boldsymbol{\beta}_0' \mathbf{x}_i - \boldsymbol{\beta}' \mathbf{x}_{i,k}}{\exp(\boldsymbol{\alpha}_0' \mathbf{u}_i + \boldsymbol{\alpha}' \mathbf{u}_{i,k})} \,\Big|\, \mathbf{p}, \boldsymbol{\gamma}\right\}, \tag{8}$$

which extends the univariate model (1) from one observation per subject to the multiple response problem, and it also allows for a generalized mixture link function

$$h(u \mid \mathbf{p}, \boldsymbol{\gamma}) = \sum_{s=1}^{d} p_s \Phi(u - \gamma_s), \tag{9}$$

where Φ is the standard normal cdf. Notice that the link (9) is obtained by marginalizing over the prior for $\delta_{i,k}$ defined by (6).

Level II (Priors for parameters). At the second stage, we define our priors for parameters. A key component of this description involves the $\boldsymbol{\beta}$ location parameter, which in applications is typically decomposed into $q \ge 1$ groups of parameters that share similar means and variances. We write $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_q')$. This type of grouping is especially applicable in the ROC context, where $\boldsymbol{\beta}$ is typically made up of parameters associated with different readers and different hospitals, and grouping is then based on anticipated similarities. Conceptually, as well as analytically, it is advantageous to model $\boldsymbol{\beta}$ by employing normal priors with hyperparameters for the mean

and variance. Although it is possible to build similar normal priors for $\boldsymbol{\alpha}$, the interpretation for the hyperparameters is less meaningful, and will complicate the details of the Gibbs sampler.

Our second stage priors are based on a selection of uniform priors for scale parameters and cutpoints, and normal priors for location parameters:

$$\begin{aligned}
(\boldsymbol{\alpha}_0, \boldsymbol{\alpha} \mid A) &\sim \text{uniform}[-A, A] \otimes \cdots \otimes \text{uniform}[-A, A] \\
(\boldsymbol{\beta}_0 \mid \mathbf{b}_0, \mathbf{W}_0) &\sim \text{N}(\mathbf{b}_0, \mathbf{W}_0) \\
(\boldsymbol{\beta}_\ell \mid b_\ell, \sigma_\ell) &\sim \text{N}((b_\ell, \ldots, b_\ell)', \sigma_\ell \mathbf{I}), \qquad \text{for } \ell = 1, \ldots, q \\
(\boldsymbol{\theta} \mid T) &\sim \text{uniform}\{-T \le \theta_1 \le \cdots \le \theta_{J-1} \le T\}, \\
&\qquad\qquad \text{where } \theta_0 = -\infty,\ \theta_J = +\infty \\
\mathbf{R} &\propto c. \qquad (10)
\end{aligned}$$

Note that in (10) we are assuming that $\mathbf{R}$ has a uniform prior on the set of proper correlation matrices. Specifically, by this we mean that we reparameterize $\mathbf{R}$ by expressing it as a vector of dimension $K^* = K(K-1)/2$ consisting of its off-diagonal elements, and then embed this vector in the hypercube $[-1, 1]^{K^*}$. The set of proper correlation matrices, when embedded this way, form a convex subset of the hypercube, and a uniform prior over this space leads to a uniform (and hence finite mass) prior for $\mathbf{R}$.

Level III (Priors for hyperparameters). The third level of the model specifies priors on the hyperparameters as follows:

$$\begin{aligned}
(\mathbf{b}_0 \mid \tau_0) &\sim \text{N}(\mathbf{0}, \tau_0 \mathbf{I}) \\
(b_\ell \mid \tau_\ell) &\sim \text{N}(0, \tau_\ell), \qquad \ell = 1, \ldots, q \\
(\mathbf{W}_0^{-1} \mid \mathbf{V}, v) &\sim \text{Wishart}(\mathbf{V}^{-1}, v) \\
(\sigma_\ell^{-1} \mid c_\ell, d_\ell) &\sim \text{gamma}(c_\ell, d_\ell), \qquad \ell = 1, \ldots, q. \qquad (11)
\end{aligned}$$

REMARKS

(i) *Constants for hyperparameters.* The parameters $(A, T, \boldsymbol{\gamma}, \mathbf{a}, \tau_0, \tau_1, \ldots, \tau_q, \mathbf{V}, v, c_1, d_1, \ldots, c_q, d_q)$ in the priors (7), (10), and (11) are fixed values that we choose in order to induce "vague" but proper priors on parameters. For example, for the means $\mathbf{b}_0$ and $\mathbf{b} = (b_1, \ldots, b_q)'$ we use large values $\tau_0 = \tau_1 = \cdots = \tau_q = 1,000$, while for $\mathbf{W}_0$, we set $\mathbf{V} = 3\mathbf{I}$ and $v = m_0 + 2$, where m_0 is the dimension for $\boldsymbol{\beta}_0$. The choice for $\mathbf{V}$ and v ensures that the prior mean for $\mathbf{W}_0$ is $\mathbf{V}$, which gives a reasonable trade-off between the noninformativeness of $\mathbf{W}_0$ and the amount of influence the data can have on the posterior for $\mathbf{W}_0$. The priors for the variances $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_q)'$ in $\boldsymbol{\beta}$ can influence the posterior to some extent, so we use a noninformative prior by choosing small values for c_ℓ and large values for d_ℓ ($c_\ell = 0.001$, $d_\ell = 1,000$). In order to elicit a flat prior for $\boldsymbol{\alpha}_0, \boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ we choose $A = 10$ and $T = 4.5$.

(ii) *Identification.* In the univariate ordinal regression model, the first, or several, coordinates of $\boldsymbol{\theta}$ are usually fixed in advance in order to avoid lack of identification. This nonidentification usually does not exist in the multivariate ordinal response case, or at least not in the case when the model is rich in scale covariates. However, without scale parameters, or if all scale parameters equal zero, the likelihood for the multiple response model remains invariant when a constant value is added to each θ_j cutpoint and to each $\delta_{i,k}$. In this case, nonidentification can be avoided by either fixing at least one θ_j cutpoint in advance, or by modeling $\delta_{i,k}$ to have one common value for each patient i (see our example in Section 4). In most ROC applications, this may not be problematic because covariate information always includes true disease status and usually some patient clinical information which can

be included as a scale effect in the model. In general, our approach is to always leave θ unconstrained, adding constraints only when the output from our Gibbs sampler indicates a need for them.

(iii) *Posterior ROC areas and estimated link functions.* We take a fully Bayesian approach to fitting HROC models, using Markov Chain Monte Carlo (MCMC) simulation of the posterior distribution of the parameters. Details are provided in the appendix. Posterior estimates of ROC curves and their functionals, such as the area under the curve, can be estimated using simulated values from the MCMC sampler. If $(\alpha_0^{(r)}, \alpha^{(r)}, \beta_0^{(r)}, \beta^{(r)}, \theta^{(r)}, \mathbf{p}^{(r)})$ denotes the r^{th} sampled value from the posterior distribution for $(\alpha_0, \alpha, \beta_0, \beta, \theta, \mathbf{p})$, then the conditional distribution function

$$\mathrm{P}\left\{Y \le r_j \mid \alpha_0^{(r)}, \alpha^{(r)}, \beta_0^{(r)}, \beta^{(r)}, \theta^{(r)}, \mathbf{p}^{(r)}\right\}, \qquad \text{for } j = 1, \ldots, J,$$

evaluated over these values can be averaged to construct a Bayesian estimate for the empirical ROC curve.

A smoothed ROC curve can also be constructed by evaluating

$$h\left\{\frac{u - \mathbf{x}_i'\beta_0^{(r)} - \mathbf{x}_{i,k}'\beta^{(r)}}{\exp(\mathbf{u}_i'\alpha_0^{(r)} - \mathbf{u}_{i,k}'\alpha^{(r)})} \;\middle|\; \mathbf{p}^{(r)}, \gamma\right\} = \sum_{s=1}^{d} p_s^{(r)} \Phi\left\{\frac{u - \mathbf{x}_i'\beta_0^{(r)} - \mathbf{x}_{i,k}'\beta^{(r)}}{\exp(\mathbf{u}_i'\alpha_0^{(r)} - \mathbf{u}_{i,k}'\alpha^{(r)})} - \gamma_s\right\}$$

over a grid of u values and then averaging over the resulting ROC curves. However, these calculations can become computationally expensive when deriving ROC areas for several different covariate values. A quicker method is to use a posterior estimate $(\widehat{\alpha}_0, \widehat{\alpha}, \widehat{\beta}_0, \widehat{\beta})$ for $(\alpha_0, \alpha, \beta_0, \beta)$ (such as the posterior mean value) and construct the ROC curve on the basis of

$$\frac{1}{R}\sum_{r=1}^{R} h\left\{\frac{u - \mathbf{x}_i'\widehat{\beta}_0 - \mathbf{x}_{i,k}'\widehat{\beta}}{\exp(\mathbf{u}_i'\widehat{\alpha}_0 - \mathbf{u}_{i,k}'\widehat{\alpha})} \;\middle|\; \mathbf{p}^{(r)}, \gamma\right\} = \widehat{h}\left\{\frac{u - \mathbf{x}_i'\widehat{\beta}_0 - \mathbf{x}_{i,k}'\widehat{\beta}}{\exp(\mathbf{u}_i'\widehat{\alpha}_0 - \mathbf{u}_{i,k}'\widehat{\alpha})}\right\},$$

where

$$\widehat{h}(u) = \sum_{s=1}^{d} \widehat{p}_s \Phi(u - \gamma_s) \tag{12}$$

is the posterior link estimate and $\widehat{\mathbf{p}} = \sum_{r=1}^{R} \mathbf{p}^{(r)}/R$.

(iv) *Goodness-of-fit tests.* A simple method for checking goodness-of-fit can be based on latent data residuals (Albert & Chib 1995). In particular, the expression for the latent variable (4) implies that

$$L_{i,k} = (M_{i,k} - \beta_0'\mathbf{x}_i - \beta'\mathbf{x}_{i,k}) \exp(-\alpha_0'\mathbf{u}_i - \alpha'\mathbf{u}_{i,k}) - \delta_{i,k}$$

are latent residuals which can be compared to a standard normal distribution for assessing goodness-of-fit. For large data sets, the disk space needed to store these residuals quickly becomes unmanageable, and instead, a more practical method can be based on comparing $\sum_i L_{i,k}^2$ to a χ^2-distribution. This method presents a goodness-of-fit test useful for identifying lack of fit in specific clusters. Alternatively, an overall goodness-of-fit assessment can be constructed by standardizing the latent residual vector $\mathbf{L}_i = (L_{i,1}, \ldots, L_{i,k_i})'$ by multiplying by $\mathbf{R}_i^{-1/2}$. If the model is adequate, the residual vectors are standard multivariate normals and $\sum_i \mathbf{L}_i'\mathbf{R}_i^{-1}\mathbf{L}_i$ can be compared to a χ^2-distribution.

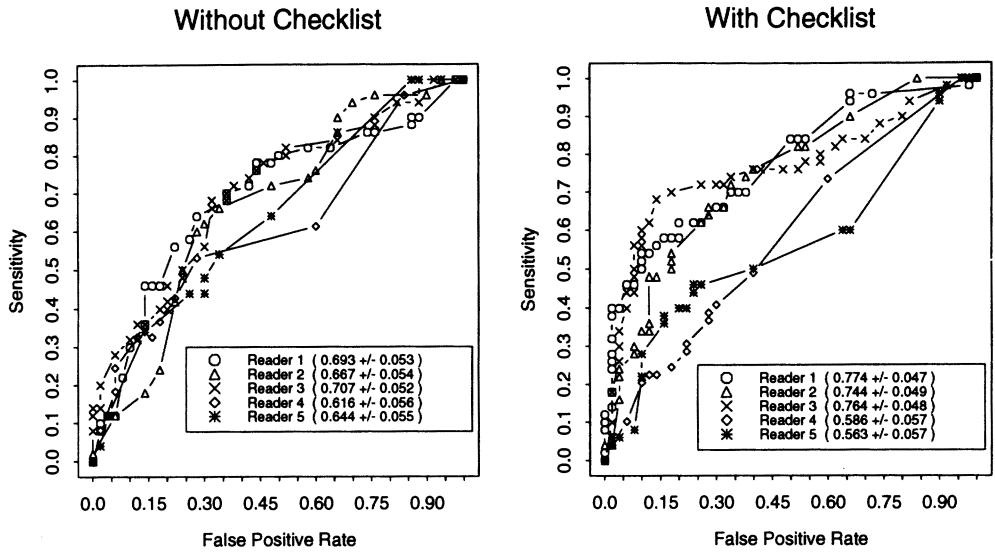

FIGURE 1: Empirical ROC curves with areas and standard errors for five readers from the enhanced prostate study (with and without the aid of the feature checklist). Standard errors are derived using normal approximations to U-statistics (DeLong, DeLong & Clarke-Pearson 1988).

4. EXAMPLE 1: ENHANCING MRI IN PROSTATE CANCER

4.1. Data.

As our first application, we analyzed data from a study involving magnetic resonance (MR) imaging of the prostate gland using endorectal coil (Seltzer *et al.* 1997). The study utilized a group of radiologists expert in prostate cancer in order to construct a list of perceptual features of MR imaging considered to be critical in identifying prostate cancer. In order to test the effectiveness of the feature checklist, a different group of radiologists was asked to elicit their overall suspicion rating on a 100-point scale to a set of 100 MR images, with and without the use of the feature checklist. This test group was made up of 5 radiologists who were considered to be less experienced in prostate MR imaging and whose clinical experience was largely based on body MR imaging. This way the group could be used to provide reasonable baseline readings. At baseline, each of the 5 test radiologists interpreted the MR scans from the same set of 100 independent patients who were equally split between Stage A/B (localized) and Stage C/D (advanced) prostate cancer (as determined by biopsy). The test group then rated the same set of MR scans in a second session (6 months later), but this time, with the aid of the perceptual feature checklist (although one reader failed to give either a baseline or enhanced reading for one of the scans).

The empirical ROC plots for each of the five readers (with and without a checklist) are given in Figure 1At the baseline, the first three readers (readers 1, 2 and 3) have a cluster of similar-looking curves with areas under the curve of 0.693, 0.667 and 0.707 (indicating reasonable accuracies for readers not specialized in prostate cancer). The remaining two readers (readers 4 and 5) also have curves that appear to be similar, but with smaller areas of 0.616 and 0.644. As we can see, when aided by the checklist, the first 3 readers realized a substantial increase in accuracy ($\chi^2_1 = 4.84$, p-value = 0.03; DeLong, DeLong & Clarke-Pearson 1988), while the remaining two readers had accuracies that decreased to some degree of significance ($\chi^2_1 = 3.18$, p-value = 0.07).

4.2. Models.

The ordinal data in this study are clustered into the $K = 10$ groups formed by the responses of the five readers at baseline and in the enhanced condition when aided by the checklist. Because the data are balanced, we can fit a reader specific parameter for each of the 10 clusters. Therefore, our models include location reader specific parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{10})'$ and scale reader specific parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{10})'$. Note that in this case, the prior for $\boldsymbol{\beta}$ is defined by

$$\begin{aligned} (\beta_k \mid b_k, \sigma_k) &\sim \mathrm{N}(b_k, \sigma_k) \\ (b_k \mid \tau_k) &\sim \mathrm{N}(0, \tau_k) \\ (\sigma_k^{-1} \mid c_k, d_k) &\sim \mathrm{Gamma}(c_k, d_k), \qquad k = 1, \ldots, 10. \end{aligned}$$

The covariates $\mathbf{u}_{i,k}$ and $\mathbf{x}_{i,k}$ for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are indicator terms that identify the cluster (i.e., which reader and in which experimental condition) and are multiplied by the true disease status ($-1/2$ for Stage A/B and $1/2$ for Stage C/D). Therefore, each covariate $\mathbf{u}_{i,k}$ and $\mathbf{x}_{i,k}$ contains a string of zeros, except for the coordinate corresponding to the cluster, which is then either $1/2$ or $-1/2$ depending on the disease status of the patient. Thus, in this parameterization, a large value for β_k shows that the reader for cluster k is accurately predicting the stage of cancer, while the value for α_k measures the reader's change in variability over the two disease groups. In addition to the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we have also included a patient specific scale parameter α_0 corresponding to the covariate information $\mathbf{u}_i$ composed of the age of the patient and their PSA (prostate-specific antigen) level.

Because of the anticipated differences between readers, we fit two different models. In Model 1, we employed a common link for all readers by assuming a common $\delta_{i,k} = \delta_i$ value for each patient i, while in Model 2, we fit 5 separate independent links for each of the 5 readers.

We followed the guideline outlined in Section 3.2 for selecting the hyperparameters in our 2 models. Also, for our support vector $\boldsymbol{\gamma}$, we used $d = 133$ equally spaced points selected from the interval $[-4, 4]$, feeling that 133 points were more than adequate based on the sample size of 1,000. Our reasoning was that any effect along this interval could not be identified with better than a $2/\sqrt{1000} = 0.06$ accuracy. Therefore, we felt little would be gained by using a grid with spacings finer than $8/0.06 = 133$ grid points. Finally, for our smoothing parameter we choose $\mathbf{a} = c\mathbf{1}$, where $c = 0.001$.

4.3. Results.

We estimated our models by invoking the Gibbs sampling scheme discussed in the appendix. In each case, we used a 100,000 iteration burn-in, which was then followed by sampling one value each 100^{th} iteration until a sample of 2,500 values was collected. The large number of burn-in iterations used, and the need to lag sampled values, is necessary to ensure low autocorrelations between parameter values. The presence of this high autocorrelation is mostly due to the method we are using for sampling the cutpoints. We use the method described in Albert & Chib (1993), which although simple to implement, can often lead to high autocorrelations in the cutpoints. For other methods for sampling cutpoints, see Cowles (1996) and Nandram & Chen (1996).

Using the values obtained from our Gibbs sampler, we computed the posterior empirical ROC areas for each of the 5 readers at baseline and with checklist. Not surprisingly, we found that Model 1 (based on a common link function) generated areas which sometimes differed as much as 15% from the observed empirical areas, while Model 2 (different links) generated areas much closer to the observed values. See Figure 2. The areas computed from our models can also be used to test whether the overall accuracy of readers was enhanced by the use of the checklist. The posterior 95% interval of the average difference in areas with and without checklist was $[-0.05, 0.05]$ from Model 1 and $[-0.05, 0.04]$ from Model 2. Thus, in both models we find no significant overall improvement.

We used the goodness-of-fit method discussed in Section 3.2 to formally compare our two models. The posterior "p-values" for the fit to each of the 10 clusters was 7×10^{-3}, 0.44, 0.22,

0.10, 10^{-4}, 0.39, 0.02, 0.25, 0.50, 0.01 from Model 1 and 0.02, 0.37, 0.25, 0.04, 0.03, 0.30, 0.38, 0.04, 0.19, 0.06 from Model 2, respectively. Thus, for clusters 1 and 5, the use of separate links in Model 2 has led to a noticeable improvement in the goodness-of-fit. The need for separate links is also apparent from Figure 3. Notice that the figure also demonstrates that our estimated links are quite different from a probit link (especially for readers 4 and 5). Clearly, a probit link would be inadequate to model such a large number of categories.

Finally, we found all scale parameters (α and α_0) to be zero, except for age in Model 2. With all scale parameters zero for clusters, we can conclude that readers are evaluating the stage A/B and stage C/D patient with equal variability.

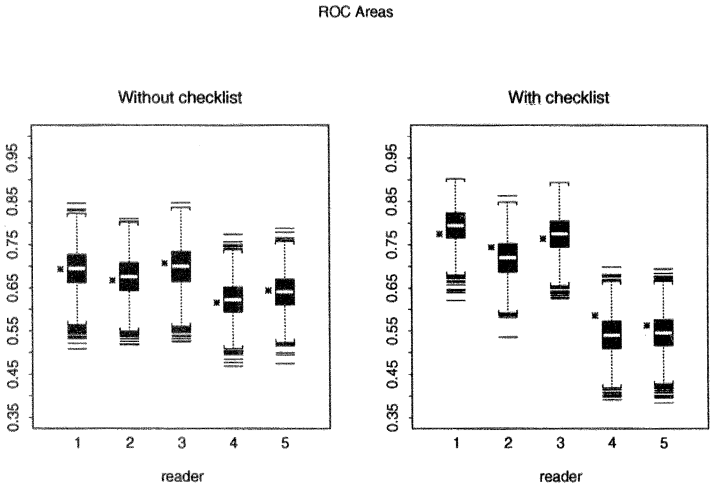

FIGURE 2: Posterior empirical ROC areas for each of the 5 readers using Model 2, with observed empirical areas indicated by a star (offset slightly to the left). Areas are derived with PSA and age covariates set at their average values.

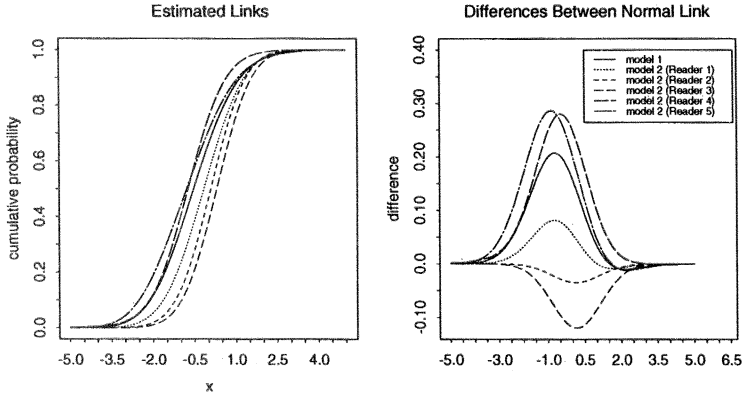

FIGURE 3: Left-hand side presents estimated link functions from Models 1 and 2 using (12). Right-hand side plots are the differences between the estimated links and the standard normal cdf. The skewed nature of some of these differences indicates a departure from the standard normal cdf beyond a shift due to location.

5. EXAMPLE 2: COMPARISON OF CT AND MRI IN STAGING HEAD AND NECK CANCER

5.1. Data.

For our second application, we analyzed data from a multi-institution, multi-reader study of computed tomography (CT) and magnetic resonance (MR) imaging in staging neck metastases (Curtin et al. 1998). The study involved a consortium of 3 participating hospitals and over 20 radiologists who read both the CT and MR images from the left and/or right-hand side of the necks of 208 patients. All CT and MR images were read by each hospital by subdividing the images amongst a group of radiologists from the participating institution. The degree of suspicion that forms the data we analyze was given on a 5-point ordinal scale representing the level of suspicion for internal abnormality (abnormal region signal) seen from the nodes on an image. The true disease status of each neck was determined through pathological examination. See Curtin et al. (1998) for more details.

All radiologists were trained in both CT and MR. This meant that the case mix for a particular radiologist usually involved both CT and MR images, but never a CT and MR image for the same neck. In some cases a radiologist read either the CT or the MR image of both the the left and right-hand side of a neck, but in these cases the readings were done independently. Radiologists were also trained to read their CT and MR images independently. Therefore, we conceptualized our pool of radiologists as a set of 38 readers, 20 readers who read CT independently of 18 readers who read MR (both the CT and MR groups contained a group of radiologists who were pooled because of their small case load). This meant that each neck in the study had a CT image which was read by 3 different CT readers from our pool of 20 (one radiologist from each hospital). The same scenario held for the MR image of a neck. Therefore, each neck in the study was read a maximum of 6 times, giving a maximum of 6 ordinal categorical responses for each neck from the $K = 6$ different clusters formed by the institution and type of reading.

5.2. Models.

Because of the large number of readers and relatively few cases per reader in this study, a hierarchical approach to analysis of reader accuracy seems quite appropriate. Our models included the location parameter $\beta = (\beta_1, \ldots, \beta_{38})'$ consisting of the 38 radiologists reader accuracy parameters, which we then grouped by experience. It was believed that some readers were much more experienced in interpreting head and neck cancer, and therefore we created four groups of parameters representing either experienced or less experienced CT readers, and either experienced or less experienced MR readers. Therefore, we partitioned β' into 4 groups $(\beta_1', \beta_2', \beta_3'\beta_4')$ with independent priors

$$\begin{aligned}(\beta_\ell \mid b_\ell, \sigma_\ell) &\sim \mathrm{N}((b_\ell, \ldots, b_\ell), \sigma_\ell \mathbf{I}) \\ (b_\ell \mid \tau_\ell) &\sim \mathrm{N}(0, \tau_l) \\ (\sigma_\ell^{-1} \mid c_\ell, d_\ell) &\sim \mathrm{Gamma}(c_\ell, d_\ell), \qquad \ell = 1, \ldots, 4.\end{aligned}$$

We also included one scale parameter for each of the 4 groups, $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$.

Somewhat like the prostate example of Section 4, our covariates $\mathbf{x}_{i,k}$ and $\mathbf{u}_{i,k}$ for β and α were indicator terms that identified the specific reader, or group (in the case of $\mathbf{u}_{i,k}$), and were multiplied by $-1/2$ or $1/2$ depending upon the true disease status of the patient. Specifically, each $\mathbf{x}_{i,k}$ was a string of zeros, except in the coordinate corresponding to the particular reader from our group of 38, which was then $1/2$ if the patient was diseased or $-1/2$ if not. Similarly, the coordinates of $\mathbf{u}_{i,k}$ were zero, except in the coordinate in which the particular reader was grouped, which was then $1/2$ or $-1/2$ depending upon the disease status of the patient. As discussed earlier, in this parameterization β measures reader accuracy in interpreting internal abnormality, while α measures the different variability over the two disease groups for the various reader groups.

TABLE 1: Posterior mean values and standard deviations for areas under ROC curves and posterior 95% intervals for between reader variability (measured by σ_ℓ); *95% posterior interval excludes zero.

		Model 1	Model 2
Areas	CT	0.79 ± 0.04	0.79 ± 0.07
	CT experts	0.80 ± 0.05	0.80 ± 0.06
	CT non-experts	0.79 ± 0.04	0.79 ± 0.07
	CT experts − CT non-experts	0.01 ± 0.03	0.01 ± 0.03
	MR	0.71 ± 0.08	0.71 ± 0.10
	MR experts	0.69 ± 0.07	0.69 ± 0.09
	MR non-experts	0.72 ± 0.08	0.71 ± 0.10
	MR experts − MR non-experts	-0.03 ± 0.03	-0.03 ± 0.04
	CT − MR	$0.08 \pm 0.02^*$	$0.08 \pm 0.02^*$
Between reader variability	CT experts	$[0.007, 0.57]$	$[0.21, 3.23]$
	CT non-experts	$[0.004, 0.19]$	$[0.15, 0.87]$
	MR experts	$[0.008, 1.69]$	$[0.28, 10.3]$
	MR non-experts	$[0.014, 0.62]$	$[0.21, 1.29]$

5.3. Results.

With a simple 5-point ordinal scale, we found it unnecessary to estimate the link function. Therefore, we set $\delta_{i,k} = 0$ in (4) to implicitly utilize a probit link. Other than this change, we ran the Gibbs sampler using the same strategy as in Section 4, employing the same guidelines for selecting hyperparameters discussed in Section 3.2. In particular, for the distributions of σ_ℓ we used a noninformative prior by setting $c_\ell = 0.001$ and $d_\ell = 1,000$. As mentioned earlier, the choice of hyperparameters for σ_ℓ can affect posterior estimates to some extent. Therefore, in order to test this sensitivity, we ran a second model (Model 2) which employed an informative prior for σ_ℓ by using $c_\ell = 1$ and $d_\ell = 1$.

From our first model (noninformative priors for σ_ℓ), the average goodness-of-fit for each of the 6 clusters had posterior "p-values" equal to 0.05, 0.17, 0.04, 0.43, 10^{-4} and 0.42, indicating a serious lack of fit for cluster 5 (MR readings at hospital 2). With such a complex problem, we can expect some lack of fit, but the small p-values are a sign that CT and MR are interpreting internal abnormality differently. This difference is also apparent in Figure 4 which contains the posterior empirical ROC areas for each of the 38 readers. As seen, the accuracy of CT is on average superior to the accuracy of MR, and when tested the posterior 95% interval for the difference was found to exclude zero (see Table 1 for more details).

Table 1 also shows that expert CT readers have roughly the same accuracy as their less experienced counterparts, while expert MR readers perform somewhat less accurately as a group than their counterparts. However, both differences have 95% posterior intervals that include zero. Table 1 also records the posterior 95% intervals for σ_ℓ, which represents a measure of reader variability for each of the 4 groups. As we see, the variability between expert readers is higher than for non-experts, and this effect is more pronounced in Model 2 due to the use of our informative prior for σ_ℓ. We also found that the scale parameters for readers were not significantly different from zero in both models, except for the parameter for the non-expert MR readers, which had a mean value of 0.54 in Model 1 and a mean value of 0.52 in Model 2. Therefore, non-expert MR readers experienced more variability in their image analysis of the diseased patients.

In general, except for σ_ℓ, the mean values for all parameters are similar in the two models, with slightly higher standard deviations seen for Model 2. With such a large data set, we found that the use of an informative prior had a minimal impact on most parameters in the model.

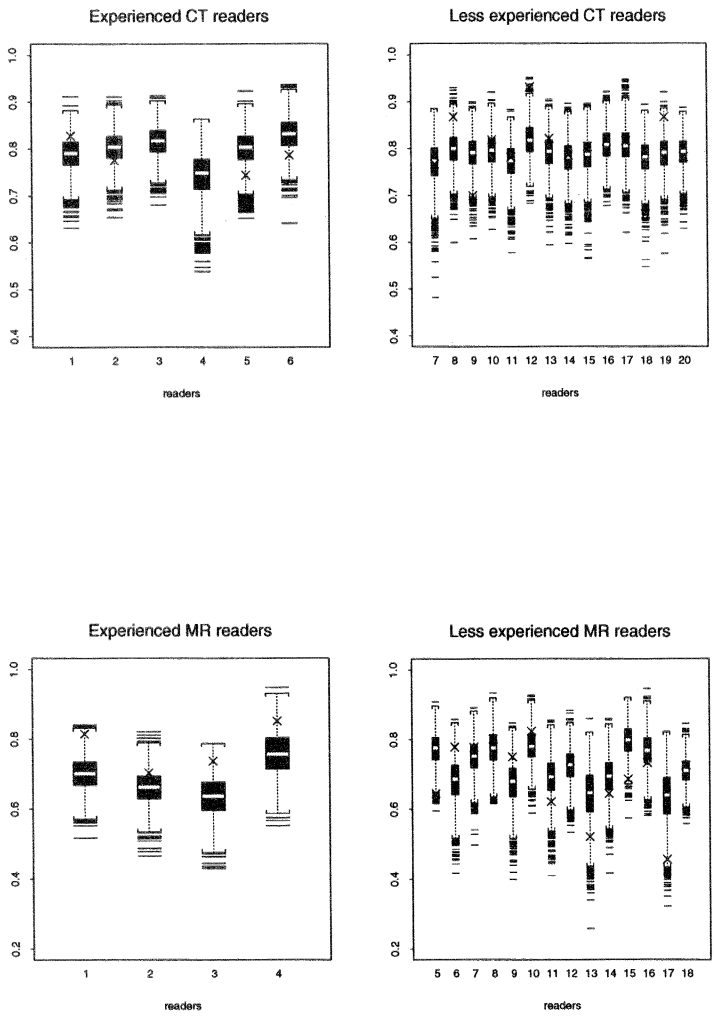

FIGURE 4: Posterior empirical ROC areas for each of the 38 readers from the head and neck study using Model 1 (noninformative priors for σ_ℓ). Observed empirical areas are indicated by a cross.

It is also important to consider the consistency of the image analysis across the 3 participating hospitals. This can be measured by considering $\widehat{\mathbf{R}}$, the posterior average of the correlation matrix $\mathbf{R}$. From Model 1, this was

$$\widehat{\mathbf{R}} = \begin{bmatrix} 1.000 & 0.506 & 0.523 & 0.470 & 0.388 & 0.274 \\ 0.506 & 1.000 & 0.591 & 0.369 & 0.464 & 0.320 \\ 0.523 & 0.591 & 1.000 & 0.335 & 0.411 & 0.235 \\ 0.470 & 0.369 & 0.335 & 1.000 & 0.263 & 0.227 \\ 0.388 & 0.464 & 0.411 & 0.263 & 1.000 & 0.435 \\ 0.274 & 0.320 & 0.235 & 0.227 & 0.435 & 1.000 \end{bmatrix},$$

where we have recorded $\widehat{\mathbf{R}}$ by ordering our 6 ordinal responses by hospitals, with CT images

followed by MR images. For example, the correlation for CT between hospital 1 and 3 was 0.523, while the correlation between CT for hospital 1 and MR for hospital 3 was 0.274. As we can see, there is a fairly high correlation within CT image analysis (upper left 3×3 matrix), while the correlation within MR analysis (lower right 3×3 matrix) is much smaller. The inconsistency in MR readings is not surprising given that readers have less experience with MR than CT in head and neck cancer.

TABLE 2: True values and posterior estimates from the simulated example. Mean values and standard deviations are based on 5,000 sampled values from the Gibbs sampler, with each value sampled every 10th iteration following a 15,000 iteration burn-in.

Type	Parameter	True Value	Mean±SD
location	$\beta_{0,1}$	2.0	2.05 ± 0.13
	$\beta_{0,2}$	1.0	1.05 ± 0.07
scale	α_1	-0.5	-0.48 ± 0.05
	α_2	-1.0	-1.06 ± 0.13
cutpoints	θ_0	$-\infty$	-
	θ_1	-2.0	-2.04 ± 0.14
	θ_2	-1.0	-0.96 ± 0.08
	θ_3	0.0	0.09 ± 0.06
	θ_4	0.5	0.62 ± 0.08
	θ_5	1.0	1.10 ± 0.09
	θ_6	2.0	2.16 ± 0.16
	θ_7	$+\infty$	-

6. SIMULATION EXAMPLE

We tested the sensitivity of our proposed model through a simulated example involving $N = 250$ three-dimensional vectors $\mathbf{Y}_i = (Y_{i,1}, Y_{i,2}, Y_{i,3})'$, with each $Y_{i,k}$ defined through the latent variable

$$M_{i,k} = \beta_0' \mathbf{x}_i + (Z_{i,k} + \delta_i) \exp(\alpha' \mathbf{u}_{i,k}) \qquad i = 1, \ldots, 250,\ k = 1, 2, 3. \tag{13}$$

In (13) we set $\delta_i = 0$ and simulated $Z_{i,k}$ from a normal distribution in order for the $Y_{i,k}$ to represent ordinal values classified by a normal latent variable. Furthermore, in order to test the sensitivity of our models in recovering correlations, we simulated the $Z_{i,k}$ according to

$$\mathbf{Z}_i = (Z_{i,1}, Z_{i,2}, Z_{i,3})' \overset{\text{i.i.d.}}{\sim} \mathrm{N}_3(\mathbf{0}, \mathbf{R}),$$

where $\mathbf{R}$ was a correlation matrix defined by

$$\mathbf{R} = \begin{bmatrix} 1.0 & -0.2 & 0.5 \\ -0.2 & 1.0 & 0.2 \\ 0.5 & 0.2 & 1.0 \end{bmatrix}.$$

The covariates in (13) were simulated independently from each other. In particular, $\mathbf{x}_i$ was a two-dimensional covariate, fixed for each value of i, with its first coordinate simulated from a uniform distribution on $\{-0.5, 0.5\}$, while its second coordinate was simulated independently from a uniform distribution on $\{-2, -1, 0, 1, 2\}$. Covariates $\mathbf{u}_{i,k}$ were also two-dimensional, with first coordinates simulated from a uniform distribution on $\{0, 1, 2\}$, and second coordinates simulated independently from a $N(0, 0.3^2)$ distribution, for each i and k. To complete the specification of our model, we chose a fixed 8-dimensional vector of cut-points for θ.

In fitting the model, we used hyperparameters selected following the guidelines of Section 3.2. We also fit a link function to test the sensitivity in recovering a probit link. This was done by selecting a support vector γ made up of $d = 133$ equally spaced points in the interval $[-4, 4]$. For the smoothing parameter for $\mathbf{p}$, we chose $\mathbf{a} = c\mathbf{1}$, with $c = 0.001$.

Results are given in Table 2 and Figure 5. Table 2 shows that the model has recovered, within reasonable accuracy, the true location, scale and cutpoint parameters, while Figure 5 presents the posterior estimate for the link function, which can be seen to be similar to the probit link function used in the simulation. The model also performed well in recovering the correlation matrix $\mathbf{R}$. The posterior mean value was

$$\hat{\mathbf{R}} = \begin{bmatrix} 1.00 & -0.22 & 0.48 \\ -0.22 & 1.00 & 0.19 \\ 0.48 & 0.19 & 1.00 \end{bmatrix},$$

which is within 0.02 of each value of $\mathbf{R}$.

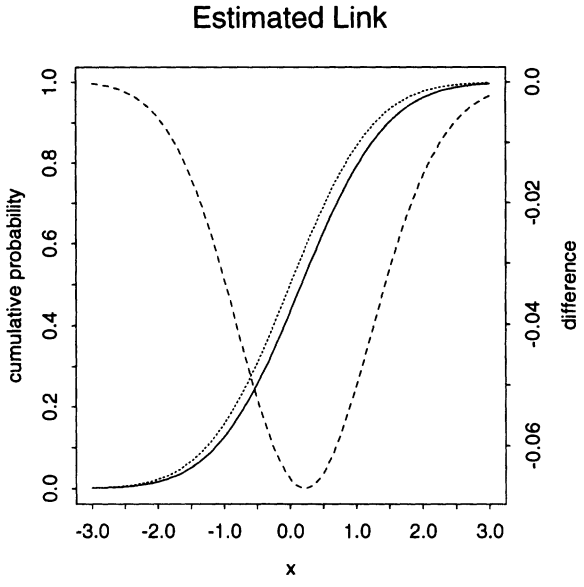

FIGURE 5: Normal cumulative distribution function (thin dashed line) and posterior estimated link function using (12) (solid line) are indicated by left-hand scale. Thick dashed line indicates the difference between the two links (right-hand scale). The observed bell-shaped difference indicates that the estimated link function is very close to a standard normal cdf, but slightly shifted by location.

7. DISCUSSION

Hierarchical modeling for ROC data is widely applicable to studies where the "effectiveness" of diagnostic procedures depends upon a broad spectrum of professionals in the delivery of everyday medical care. Example 2, of Section 5, involved data from such a study and illustrated some of the potential uses of HROC models. In particular, the effectiveness of CT and MR in determining the extent of head and neck cancer was quantified in terms of their overall accuracies when

used by either an "expert" or "non-expert" radiologist. Effectiveness was also measured by the consistency in which the two procedures were used across differing institutions and radiologists. Furthermore, the analysis provided an assessment of the components of variation between the broad group of readers used in the study, and an estimate for the variability in image interpretation depending upon the true disease status of the patient. We note that an HROC analysis can be applied to other types of study designs than the one considered here. In the experimental design used in our example, each scan was assigned to a randomly selected subgroup of radiologists from a nonrandomly selected pool of radiologists who participated in the study. However, the same methodology can be applied to studies in which each image is interpreted by each of a large random sample of radiologists, who may possibly be stratified by training, experience and practice. Designs of this type are currently being employed using a two-stage sampling format in which both radiologists and institutions are randomly sampled as a method for studying the effectiveness of diagnostic procedures in a larger population (Beam 1995; Gatsonis 1995; Beam, Layde & Sullivan, 1996).

Example 1 presented an analysis involving adaptive link estimation in which the response was measured on a 100-point scale. As we found, a simplistic model (Model 1) based on a common link for all readers led to accuracy values that were as different as 15% from observed empirical accuracies—far more than the expected differences of 3–5% often seen in practice. Accuracy values closer to the observed values were found by extending the model to include separate link functions for each of the readers (Model 2). In neither model was a probit link function found to be appropriate (see Figure 3. We were also able to use formal goodness-of-fit χ^2-tests for identifying lack of fit in specific clusters and for testing the suitability of the multiple-link model.

The applications presented in this paper were confined to analyses where the response variable was derived from the interpreter and in which the true disease status was introduced as a covariate. The accuracy of the diagnostic procedure used by the interpreter was measured by the area under the ROC curve. Although the area is the only ROC summary measure used, the methodology discussed in the paper can be applied to other measures, such as partial areas.

Another important application of the methodology applies to models in which the disease status is the response variable, and in which the interpreters response is introduced as a covariate. This variation is useful because it presents a model for studying the positive and negative predictive values for a test. This can be especially useful in diagnostic studies involving metastatic involvement of cancer, where the disease status is the stage of cancer, described in terms of some ordinal categorical scale (such as the TNM scale, for example). A point to keep in mind in this approach, is that the disease status is fixed for each patient i, while the interpreters' responses may of course be different from one reader to the next. Therefore, the classification for $Y_{i,k}$, the true disease status of patient i for reader k, is now expressed as

$$Y_{i,1} = \cdots = Y_{i,k_i} = r_j \qquad \text{if and only if} \qquad \theta_{j-1} < M_{i,1}, \ldots, M_{i,k_i} \leq \theta_j,$$

where $M_{i,k}$ is a mixture variable that will depend upon covariate information, including the readers response. Because of the perfect correlation across $Y_{i,k}$, it may sometimes be necessary to fix the correlation matrix $\mathbf{R}$ at some value, like $\mathbf{I}$, to avoid the possibility of estimating a degenerate matrix. Except for this caveat, the MCMC approach follows the same strategy outlined in our examples and in the appendix.

We also note that HROC modeling is applicable to meta-analysis of diagnostic test evaluations, if the original ROC data are available from each study. This is true even if the different studies involve variable numbers of readers with different levels of experience. Furthermore, continuous covariate information about the readers, such as the number of images read per year, can also be incorporated into the HROC models. Although this paper has implicitly only considered categorical reader specific information by the method of grouping the means and variances of β, the normal priors used for β could be extended, for example, by allowing the means to depend upon continuous covariates.

APPENDIX: GIBBS SAMPLING DETAILS

For notational convenience, we will adopt the following system. For parameters indexed by i, we use a symbol subscripted by an i to denote the vector (or matrix) formed by collecting parameters over the k_i different readings for subject i. Furthermore, when there is no ambiguity, we drop the subscript of i to indicate aggregation over subjects. Thus, for example, $\boldsymbol{\delta}_i$ denotes the k_i-dimensional vector of individual parameters $(\delta_{i,1}, \delta_{i,2}, \ldots, \delta_{i,k_i})'$ for subject i, while $\boldsymbol{\delta}$ denotes the stacked $N^* \times 1$ vector $(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_N)'$, where $N^* = k_1 + \cdots + k_N$ is the total number of observations. Covariate matrices are defined somewhat similarly. We let

$$\mathbf{X}_i = \begin{bmatrix} (\mathbf{x}_i', \mathbf{x}_{i,1}') \\ \vdots \\ (\mathbf{x}_{k_i}', \mathbf{x}_{i,k_i}') \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_N \end{bmatrix}.$$

Thus, if m is the combined dimension of $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}$, then $\mathbf{X}_i$ is the $k_i \times m$ matrix of location covariates for subject i, while $\mathbf{X}$ is the $N^* \times m$ location covariate matrix.

The Gibbs sampler we use is based on the data augmentation approach described in Albert & Chib (1993), which involves augmenting the posterior to include the latent variables $\mathbf{M}$. In particular, to sample values from our posterior, we use the Gibbs sampler to sample

$$(\mathbf{M}, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{R}, \mathbf{p}, \mathbf{b}_0, \mathbf{b}, \mathbf{W}_0, \boldsymbol{\sigma} \mid \mathbf{Y}),$$

and for this we will need to sample the full conditionals of the following parameters:

$$\begin{aligned}
\mathbf{M} &\mid \boldsymbol{\alpha}_0, \boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{R}, \mathbf{Y} \\
(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}) &\mid \mathbf{M}, \boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\delta}, \mathbf{R} \\
(\boldsymbol{\beta}_0, \boldsymbol{\beta}) &\mid \mathbf{M}, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}, \boldsymbol{\delta}, \mathbf{R}, \mathbf{b}_0, \mathbf{b}, \mathbf{W}_0, \boldsymbol{\sigma} \\
\boldsymbol{\theta} &\mid \mathbf{M}, \mathbf{Y} \\
\boldsymbol{\delta} &\mid \mathbf{M}, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}, \mathbf{p}, \mathbf{R} \\
\mathbf{R} &\mid \mathbf{M}, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\delta} \\
\mathbf{p} &\mid \boldsymbol{\delta} \\
(\mathbf{b}_0, \mathbf{b}) &\mid \boldsymbol{\beta}_0, \boldsymbol{\beta}, \mathbf{W}_0, \boldsymbol{\iota} \\
\mathbf{W}_0 &\mid \boldsymbol{\beta}_0, \mathbf{b}_0 \\
\boldsymbol{\sigma} &\mid \boldsymbol{\beta}, \mathbf{b}.
\end{aligned}$$

The details for simulating from each of these conditionals are as follows.

- The $\mathbf{M}_i$ are conditionally independent multivariate normals constrained to lie in the k_i-dimensional rectangle defined by $\boldsymbol{\theta}$ and $\mathbf{Y}_i$ (see relationship (3)). Various standard methods exists for drawing these values. See Geweke (1991).
- The density for the conditional distribution of $(\boldsymbol{\alpha}_0, \boldsymbol{\alpha})$ is proportional to

$$\begin{aligned}
&\prod_{i=1}^{N} |2\pi\boldsymbol{\Sigma}_i|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\delta}_i - \widehat{\mathbf{M}}_i)'\mathbf{R}_i^{-1}(\boldsymbol{\delta}_i - \widehat{\mathbf{M}}_i)\right\} \\
&\qquad = c \prod_{i=1}^{N} \exp\left(-k_i\boldsymbol{\alpha}_0'\mathbf{u}_i - \sum_{k=1}^{k_i} \boldsymbol{\alpha}'\mathbf{u}_{i,k} + \boldsymbol{\delta}_i'\mathbf{R}_i^{-1}\widehat{\mathbf{M}}_i - \frac{1}{2}\widehat{\mathbf{M}}_i'\mathbf{R}_i^{-1}\widehat{\mathbf{M}}_i\right), \qquad (14)
\end{aligned}$$

where $\widehat{\mathbf{M}}_i = \boldsymbol{\Lambda}_i^{-1}(\mathbf{M}_i - \mathbf{X}_i\boldsymbol{\beta}_*)$.

- The full conditional distribution for $(\boldsymbol{\beta}_0, \boldsymbol{\beta})$ is,

$$(\boldsymbol{\beta}_0, \boldsymbol{\beta} \mid \mathbf{M}, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}, \boldsymbol{\delta}, \mathbf{R}, \mathbf{b}_0, \mathbf{b}, \mathbf{W}_0, \boldsymbol{\sigma}) \sim N\left(\mathbf{W}^* \left(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\widetilde{\mathbf{M}} + \mathbf{W}^{-1}\mathbf{b}_*\right), \mathbf{W}^*\right),$$

where $\widetilde{\mathbf{M}}_i = \mathbf{M}_i - \mathbf{D}_i\boldsymbol{\lambda}_i$, $\mathbf{W}^* = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X} + \mathbf{W}^{-1})^{-1}$, $\mathbf{W} = \text{diag}(\mathbf{W}_0, \sigma_1\mathbf{I}, \ldots, \sigma_q\mathbf{I})$, $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_N)$ and $\mathbf{b}_*' = (\mathbf{b}_0', \mathbf{b}')$. Following our notational convention, we form the $N^* \times 1$ vector $\widetilde{\mathbf{M}}$ by stacking $\widetilde{\mathbf{M}}_i$ over subjects i.

- We sample the cutpoints $(\boldsymbol{\theta} \mid \mathbf{M}, \mathbf{Y})$ using the method described in Albert & Chib (1993). Other methods for sampling $\boldsymbol{\theta}$ can be based on Cowles (1996) and Nandram & Chen (1996).

- Let $\boldsymbol{\delta}_i^{(k)}$ denote the subvector of $\boldsymbol{\delta}_i$ after removing the k^{th} coordinate. Also, for each coordinate γ_s in the support vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d)'$, define

$$\boldsymbol{\nu}_{i,k}(s) = \left(\delta_{i,1}, \ldots, \delta_{i,(k-1)}, \gamma_s, \delta_{i,(k+1)}, \ldots, \delta_{i,k_i}\right)'.$$

Then

$$P(\delta_{i,k} = \gamma_s \mid \mathbf{M}_i, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\delta}_i^{(k)}, \mathbf{p}, \mathbf{R}) = \frac{p_s\, g_{i,k,s}}{\sum_{s=1}^{d} p_s\, g_{i,k,s}},$$

where

$$g_{i,k,s} = \exp\left\{-\frac{1}{2}\boldsymbol{\nu}_{i,k}(s)'\mathbf{R}_i^{-1}(\boldsymbol{\nu}_{i,k}(s) - 2\widehat{\mathbf{M}}_i)\right\}, \qquad \text{for } s = 1, \ldots, d.$$

- With a uniform prior for $\mathbf{R}$, the conditional density for $\mathbf{R}$ is proportional to

$$\prod_{i=1}^{N} |2\pi\mathbf{R}_i|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{M}_i - \boldsymbol{\mu}_i)'\boldsymbol{\Lambda}_i^{-1}\mathbf{R}_i^{-1}\boldsymbol{\Lambda}_i^{-1}(\mathbf{M}_i - \boldsymbol{\mu}_i)\right\}. \tag{15}$$

- By the conjugacy of the Dirichlet distribution to multinomial sampling

$$(\mathbf{p} \mid \boldsymbol{\delta}) \sim \text{Dirichlet}_d(a_1^*, \ldots, a_d^*),$$

where $a_s^* = a_s + \#\{\delta_{i,k} = \gamma_s\}$ and $\#$ is the cardinality of a set.

- By conjugacy,

$$\begin{aligned}(\mathbf{b}_0 \mid \boldsymbol{\beta}_0, \mathbf{W}_0) &\sim \text{N}\left(\mathbf{B}_0^*\mathbf{W}_0^{-1}\boldsymbol{\beta}_0, \mathbf{B}_0^*\right)\\ (b_\ell \mid \boldsymbol{\beta}_\ell, \sigma_\ell) &\sim \text{N}\left(\frac{\tau_\ell^*}{\sigma_\ell}\sum_{t=1}^{m_\ell}\beta_{\ell,t}, \tau_\ell^*\right) \qquad \text{for } \ell = 1, \ldots, q\end{aligned}$$

where $\mathbf{B}_0^* = (\mathbf{W}_0^{-1} + \tau_0^{-1}\mathbf{I})^{-1}$ and $\tau_\ell^* = (1/\tau_\ell + m_\ell/\sigma_\ell)^{-1}$.

- By conjugacy,

$$(\mathbf{W}_0^{-1} \mid \boldsymbol{\beta}_0, \mathbf{b}_0) \sim \text{Wishart}\left(\{\mathbf{V} + (\boldsymbol{\beta}_0 - \mathbf{b}_0)(\boldsymbol{\beta}_0 - \mathbf{b}_0)'\}^{-1}, v + 1\right).$$

- By conjugacy,

$$(\sigma_\ell^{-1} \mid \boldsymbol{\beta}_\ell, b_\ell) \sim \text{Gamma}(c_\ell^*, d_\ell^*), \qquad \text{for } \ell = 1, \ldots, q,$$

where $c_\ell^* = c_\ell + m_\ell/2$ and $d_\ell^* = \{1/d_\ell + \sum_{t=1}^{m_\ell}(\beta_{\ell,t} - b_\ell)^2/2\}^{-1}$.

All draws described above are straightforward, except for (14) and (15), which we completed using random walk Metropolis–Hastings. In employing Metropolis–Hastings in (15), we moved from one correlation matrix $\mathbf{R}^{m-1}$ to a new matrix $\mathbf{R}^m$ by the random walk $\mathbf{R}^m = \mathbf{R}^{m-1} + \varepsilon\mathbf{S}$, where $\varepsilon > 0$ was some small fixed number and $\mathbf{S}$ a random symmetric matrix generated by placing zeros along its diagonal and by placing random values $s_{i,j}$ on the off-diagonal sampled from the uniform distribution on the unit sphere in $\mathbb{R}^{K^*}$, where $K^* = K(K-1)/2$ (for $K > 2$ we generate $s_{i,j}$ using independent standard normals standardized by their $\mathcal{L}_2$ norm). For other methods of sampling correlation matrices, see Chib & Greenberg (1998).

ACKNOWLEDGEMENTS

This research was supported in part by grant funds from the Natural Sciences and Engineering Research Council of Canada. The authors are very grateful to the helpful and constructive comments of the reviewers of an earlier draft of this paper.

REFERENCES

J. H. Albert & S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88, 669–679.

J. H. Albert & S. Chib (1995). Bayesian residual analysis for binary response regression models. *Biometrika*, 82, 747–759.

C. A. Beam (1995). Random-effects models in the receiver operating characteristic curve-based assessment of the effectiveness of diagnostic imaging technology: concepts, approaches and issues. *Academic Radiology*, 2 *Supplement*, S5–S13.

C. A. Beam, P. M. Layde & D. C. Sullivan (1996). Variability in the interpretation of screening mammograms by US radiologists. *Archives of Internal Medicine*, 156, 209–213.

S. Chib & E. Greenberg (1998). Analysis of multivariate probit models. *Biometrika*, 85, 347–361.

M. K. Cowles (1996). Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing*, 6, 101–111.

M. K. Cowles, B. P. Carlin & J. E. Connett (1996). Bayesian Tobit modeling of longitudinal ordinal clinical trial compliance data with nonignorable missingness. *Journal of the American Statistical Association*, 91, 86–98.

H. D. Curtin, H. Ishwaran, A. A. Mancuso, R. W. Dalley, D. J. Caudry & B. J. McNeil (1998). Comparison of CT and MR in staging of neck metastases. *Radiology*, 207, 123–130.

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

A. Erkanli, D. Stangl & P. Müller (1993). A Bayesian analysis of ordinal data using mixtures. *ASA Proceedings of the Section on Bayesian Statistical Science*, 51–56.

C. A. Gatsonis (1995). Random-effects models for diagnostic accuracy data. *Academic Radiology*, 2 *Supplement*, S14–S21.

J. Geweke (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints. In *Computing Science and Statistics, Volume 23, Proceedings of the 23rd Symposium on the Interface* (E. M. Keramidas and S. M. Kaufman, eds.), 571–578.

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

D. A. Harville & R. W. Mee (1984). A mixed-model procedure for analyzing ordered categorical data. *Biometrics*, 40, 393–408.

M. Hellmich, K. R. Abrams, D. R. Jones & P. C. Lambert (1998). A Bayesian approach to a general regression model for ROC curves. *Medical Decision Making*, 18, 436–443.

V. E. Johnson (1996). On Bayesian analysis of multirater ordinal data: an application to automated essay grading. *Journal of the American Statistical Association*, 91, 42–51.

P. McCullagh (1979). *PLUM: An Interactive Computer Package for Analyzing Ordinal Data*. Department of Statistics, University of Chicago, Chicago, IL.

P. McCullagh (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society Series B*, 42, 109–142.

B. Nandram & M-H. Chen (1996). Reparameterizing the generalized linear model to accelerate Gibbs sampler convergence. *Journal of Statistical Computation and Simulation*, 54, 129–144,

N. Obuchowski (1995). Multi-reader multi-modality ROC studies: hypothesis testing and sample size estimation using an ANOVA approach with dependent observations. *Academic Radiology*, 2 *Supplement*, S22–S29.

M. S. Pepe (1997). A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika*, 84, 595–608.

H. E. Rockette, D. Gur & C. E. Metz (1992). The use of continuous and discrete confidence judgments in receiver operating characteristic studies of diagnostic imaging techniques. *Statistics in Radiology*, 27, 169–172.

D. O. Stram, L. J. Wei & J. H. Ware (1988). Analysis of repeated ordered categorical outcomes with possibly missing observations and time-dependent covariates. *Journal of the American Statistical Association*, 83, 631–637.

S. E. Seltzer, D. J. Getty, C. M. C. Tempany, R. M. Pickett, M. D. Schnall, B. J. McNeil & J. A. Swets (1997). Staging prostate cancer with MR imaging: a combined radiologist-computer system. *Radiology*, 202, 219–226.

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

A. T. Toledano (1993). *Generalized Estimating Equations for Repeated Ordinal Categorical Data, With Applications to Diagnostic Medicine*. Doctoral dissertation, Harvard School of Public Health, Cambridge, MA.

A. T. Toledano & C. A. Gatsonis (1995). Regression analysis of correlated receiver operating characteristic data. *Academic Radiology*, 2 *(Supplement 1)*, 14–21.

A. N. Tosteson & C. B. Begg (1988). A general regression methodology for ROC curve estimation. *Medical Decision Making*, 8, 204–215.

J. S. Uebersax (1993). Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of the American Statistical Association*, 88, 421–427.

J. S. Uebersax & W. M. Grove (1993). A latent trait finite mixture model for the analysis of rating agreement. *Biometrics*, 49, 823–835.

J. M. Williamson, K. Kim & S. R. Lipsitz (1995). Analyzing bivariate ordinal data using a global odds ratio. *Journal of the American Statistical Association*, 90, 1432–1437

Received 17 December 1998
Accepted 24 March 2000

Hemant ISHWARAN: ishwaran@bio.ri.ccf.org
Department of Biostatistics and Epidemiology
Cleveland Clinic Foundation, Cleveland, OH 44195, USA

Constantine A. GATSONIS: gatsonis@stat.brown.edu
Center for Statistical Sciences, Brown University
Providence, RI 02912, USA