Bayesian hierarchical model for the estimation of proper receiver operating characteristic curves using stochastic ordering

Eun Jin Jang^a, Dal Ho Kim^{1,b}

^aDepartment of Information Statistics, Andong National University, Korea;

^bDepartment of Statistics, Kyungpook National University, Korea

---

Abstract

Diagnostic tests in medical fields detect or diagnose a disease with results measured by continuous or discrete ordinal data. The performance of a diagnostic test is summarized using the receiver operating characteristic (ROC) curve and the area under the curve (AUC). The diagnostic test is considered clinically useful if the outcomes in actually-positive cases are higher than actually-negative cases and the ROC curve is concave. In this study, we apply the stochastic ordering method in a Bayesian hierarchical model to estimate the proper ROC curve and AUC when the diagnostic test results are measured in discrete ordinal data. We compare the conventional binormal model and binormal model under stochastic ordering. The simulation results and real data analysis for breast cancer indicate that the binormal model under stochastic ordering can be used to estimate the proper ROC curve with a small bias even though the sample sizes were small or the sample size of actually-negative cases varied from actually-positive cases. Therefore, it is appropriate to consider the binormal model under stochastic ordering in the presence of large differences for a sample size between actually-negative and actually-positive groups.

Keywords: area under the curve, Bayesian hierarchical model, discrete ordinal data, receiver operating characteristic curve, stochastic ordering

---

1. Introduction

Diagnostic tests in medical fields detect or diagnose a disease with results measured by continuous or discrete ordinal data. The receiver operating characteristic (ROC) analysis is widely used to assess the diagnostic test performance and often presented as the ROC curve. It graphically represents the relationship between false positive and true positive rates. The false positive rate represents “1 – specificity” which indicates the probability that a truly non-diseased individual displays a positive test result, and the true positive rate represents “sensitivity” indicating the probability that a diseased individual will show a positive test result. The area under the curve (AUC) is often used to measure the accuracy of a ROC curve.

Numerous ROC curve estimation methods for continuous or discrete ordinal data have been proposed using a parametric, semiparametric, and nonparametric approach based on the frequentist or Bayesian method (Gonçalves *et al.*, 2014). The most common model is a conventional binormal model

---

¹ Corresponding author: Department of Statistics, Kyungpook National University, 80 Daehakro, Bukgu, Daegu 41566, Korea. E-mail: dalkim@knu.ac.kr

which the distributions of the decision-variables of actually-negative and actually-positive populations are assumed to be the normal distribution. The conventional binormal model for discrete ordinal data can be estimated using a maximum likelihood method based on frequentist methods (Dorfman and Alf, 1969; Swets, 1986; Hanley, 1988; Metz, 1989; Metz *et al.*, 1998) or Bayesian methods (Peng and Hall, 1996; Ishwaran and Gatsonis, 2000; Johnson and Johnson, 2006; Wang *et al.*, 2007).

If the diagnostic test is effective, the ROC curve should be concave (Dorfman *et al.*, 1996). The concave ROC curve is located above the main diagonal, which implies that $\text{ROC}(u) > u$ for $0 \leq u \leq 1$. It is referred to as proper if these conditions are satisfied and the non-concave curve with a hook is termed improper (Bandos *et al.*, 2017). Based on their experiences, Pesce *et al.* (2010) outlined the increased probability of a hook when the total number of sample cases is small, the ratio of actually-negative to actually-positive cases is far from 1, the population ROC curve is strongly skewed, or when the distribution of operating point along the population ROC curve is very uneven. Bandos *et al.* (2017) investigated the effect of severe improperness of the fitted binormal ROC curves on the AUC estimates. They generated simulated data from an actually proper ROC curve and fitted the binormal model using a maximum likelihood approach (Dorfman and Alf, 1969). However, they only considered cases where actually-negative and actually-positive groups were of the same sample size.

To estimate the proper ROC curve for discrete ordinal data, Metz and Pan (1999) proposed a proper binormal model and a new algorithm using the monotonic transformation of the likelihood ratio. In the frequentist method, the bi-gamma (Dorfman *et al.*, 1996; Hughes and Bhattacharya, 2013) and the bi-beta model (Mossman and Peng, 2016) were proposed. Recently, Nandram and Peiris (2018) developed a robust Bayesian model using stochastic ordering to obtain proper ROC curves. Additionally, Hwang and Chen (2015) applied nonparametric Bayesian methods to estimate ROC curve/AUC under stochastic ordering (Gelfand and Kottas, 2001).

Pisano *et al.* (2005) analyzed the diagnostic performance of breast cancer screening mammography using the ordinal discrete scale including the 7-point malignancy score as well as breast imaging reporting and data system (BIRADS) score. The study detected 355 cases of breast cancer from 42,745 subjects through a film mammography screening test. The estimated ROC curves of film mammography using a conventional binormal model were improper and had a hook. In the case of low-prevalence disease, the number of people who actually have the disease, based on screening diagnostic tests is generally lower compared to those without a disease. Therefore, it is important to assess the performance of the conventional binormal model and compare with the model to estimate the proper ROC curve in cases of large differences in sample size.

In this study, we apply the stochastic ordering method in a Bayesian hierarchical model to estimate of the proper ROC curve and AUC when diagnostic test results are measured with discrete ordinal data. We describe the conventional binormal model and binormal model under stochastic ordering in Section 2, and compare these two models for various sample sizes using simulations in Section 3. In Section 4, we apply two models on breast cancer data. Finally, the conclusions are discussed in Section 5.

2. Bayesian hierarchical model for ROC curve

We consider K discrete ordinal categories in actually-negative and actually-positive population. Let n_{1j} and n_{2j} denote the observed frequency in the j^{th} category, $j = 1, \dots, K$, in actually-negative cases and actually-positive cases, respectively. Then, $\mathbf{n}_1 = (n_{11}, \dots, n_{1K})^T$ indicates the vector of the observed frequencies in actually-negative cases and $n_1 = \sum_{j=1}^K n_{1j}$ is the total number of cases in actually-negative cases. Similarly, $\mathbf{n}_2 = (n_{21}, \dots, n_{2K})^T$ and $n_2 = \sum_{j=1}^K n_{2j}$ are defined in the actually-

positive cases.

Suppose that the probability of a response in category j in the actually-negative population is p_{1j} and the probability of a response in category j in the actually-positive population is p_{2j} . Then the categorical data in the actually-negative population follow multinomial distributions with probabilities $\mathbf{p}_1 = (p_{11}, \dots, p_{1K})^T$, and the categorical data in the actually-positive population follow multinomial distribution with probabilities $\mathbf{p}_2 = (p_{21}, \dots, p_{2K})^T$ (Bhattacharya and Nandram, 1996). Therefore, the joint likelihood function is expressed as:

$$p(\mathbf{n}_1, \mathbf{n}_2 | \mathbf{p}_1, \mathbf{p}_2) = \frac{n_1! n_2!}{\prod_{j=1}^K n_{1j}! n_{2j}!} \prod_{j=1}^K p_{1j}^{n_{1j}} p_{2j}^{n_{2j}}.$$

Suppose that the latent decision-variable axis is partitioned into K categories by $K - 1$ boundaries $c_1, c_2, \dots, c_{K-1}$ in the ROC model with K ordinal categories and the vector of boundaries is $\mathbf{c} = (c_1, \dots, c_{K-1})$. The beginning and ending of the decision-variable axis are assumed as $c_0 \equiv -\infty$ and $c_K \equiv \infty$, respectively (Metz *et al.*, 1998; Metz and Pan, 1999).

Let F_1 be the cumulative distribution function (CDF) in the actually-negative population and F_2 be the CDF in the actually-positive population. Therefore, the probabilities of a response in category j in the actually-negative population and the actually-positive population are

$$p_{1j} = F_1(c_j) - F_1(c_{j-1}), \quad p_{2j} = F_2(c_j) - F_2(c_{j-1}). \quad (2.1)$$

Assume the decision-variables are distributed $N(0, 1)$ in the actually-negative population and $N(\mu, \sigma^2)$ in the actually-positive population (Nandram and Peiris, 2018). The hyperprior for μ is assumed to be Cauchy prior (Gelman *et al.*, 2008) and the hyperprior for σ^2 is assumed as shrinkage prior to avoid difficulties associated with improper priors of the form $\pi(\sigma^2) \propto 1/\sigma^2$ (Gelman, 2006; Nandram *et al.*, 2013) as:

$$\pi(\mu) = \frac{1}{\pi(1 + \mu^2)}, \quad \pi(\sigma^2) = \frac{1}{(1 + \sigma^2)^2}.$$

Finally, we assume the standard logistic distribution with location 0 and scale 1 as a prior for the boundaries $\mathbf{c}$ (Nandram and Peiris, 2018),

$$c_1, c_2, \dots, c_{K-1} \stackrel{iid}{\sim} \text{logistic}(0, 1), \quad -\infty < c_1 < c_2 < \dots < c_{K-1} < \infty,$$

and the joint prior distribution for $\mathbf{c}$ is represented as

$$\pi(\mathbf{c}) = (K - 1)! \prod_{j=1}^{K-1} \frac{e^{c_j}}{(1 + e^{c_j})^2}, \quad -\infty < c_1 < c_2 < \dots < c_{K-1} < \infty.$$

2.1. Conventional binormal model

In the conventional binormal model without considering the stochastic ordering, the two CDFs are

$$F_1(c_j) = \Phi(c_j), \quad F_2(c_j) = \Phi\left(\frac{c_j - \mu}{\sigma}\right),$$

and the probabilities of a response in category j in actually-negative and actually-positive population in (2.1) are

$$p_{1j} = \Phi(c_j) - \Phi(c_{j-1}), \quad p_{2j} = \Phi\left(\frac{c_j - \mu}{\sigma}\right) - \Phi\left(\frac{c_{j-1} - \mu}{\sigma}\right).$$

Thus, the joint posterior density of $\nu, \sigma^2, \mathbf{c}$ using Bayes' theorem is given by

$$\begin{aligned} \pi(\mu, \sigma^2, \mathbf{c} | \mathbf{n}_1, \mathbf{n}_2) &\propto \prod_{j=1}^K [\Phi(c_j) - \Phi(c_{j-1})]^{n_{1j}} \left[\Phi\left(\frac{c_j - \mu}{\sigma}\right) - \Phi\left(\frac{c_{j-1} - \mu}{\sigma}\right) \right]^{n_{2j}} \\ &\quad \times \frac{1}{\pi(1 + \mu^2)} \frac{1}{(1 + \sigma^2)^2} \prod_{j=1}^{K-1} \frac{e^{c_j}}{(1 + e^{c_j})^2}. \end{aligned}$$

The conditional posterior distributions in the conventional binormal model are calculated similar to the binormal model under stochastic ordering and we use the grid method for the Markov chain Monte Carlo (MCMC) computations.

2.2. Binormal model under stochastic ordering

The stochastic order of two populations is defined as: F_2 is stochastically larger than F_1 if $F_1(u) \geq F_2(u)$ for all u (Gelfand and Kottas, 2001; Hwang and Chen, 2015).

The ROC curve is proper when $\text{ROC}(u) > u$ for $0 \leq u \leq 1$ if and only if $F_1(u) \geq F_2(u)$ (Hanson *et al.*, 2008). Therefore, two CDFs can be defined to estimate the proper ROC curve using the stochastic ordering method (Nandram and Peiris, 2018),

$$F_1(c_j) = \Phi(c_j), \quad F_2(c_j) = \Phi(c_j) \Phi\left(\frac{c_j - \mu}{\sigma}\right),$$

where $\Phi(z)$ is the standard normal CDF and the probabilities of a response in category j in actually-negative and actually-positive populations in (2.1) and can be rewritten as:

$$p_{1j} = \Phi(c_j) - \Phi(c_{j-1}), \quad p_{2j} = \Phi(c_j) \Phi\left(\frac{c_j - \mu}{\sigma}\right) - \Phi(c_{j-1}) \Phi\left(\frac{c_{j-1} - \mu}{\sigma}\right).$$

Thus, the joint posterior density of $\mu, \sigma^2, \mathbf{c}$ using Bayes' theorem is given by

$$\begin{aligned} \pi(\mu, \sigma^2, \mathbf{c} | \mathbf{n}_1, \mathbf{n}_2) &\propto \prod_{j=1}^K [\Phi(c_j) - \Phi(c_{j-1})]^{n_{1j}} \left[\Phi(c_j) \Phi\left(\frac{c_j - \mu}{\sigma}\right) - \Phi(c_{j-1}) \Phi\left(\frac{c_{j-1} - \mu}{\sigma}\right) \right]^{n_{2j}} \\ &\quad \times \frac{1}{\pi(1 + \mu^2)} \frac{1}{(1 + \sigma^2)^2} \prod_{j=1}^{K-1} \frac{e^{c_j}}{(1 + e^{c_j})^2}. \end{aligned}$$

We use the Griddy Gibbs sampler to generate samples from the posterior density because the conditional posterior distributions are not standard form (Lee *et al.*, 2017; Molina *et al.*, 2014; Nandram *et al.*, 2011; Nandram and Yin, 2016). For the grid method, we divide 100 intervals between 0 and 1, and calculate 100 mid-points for each sub-interval. We then calculate the values according to the mid-points in the conditional posterior density, and divide by the sum of the calculated values to convert it to an approximate probability mass function. To draw a random sample, we choose one of the

sub-intervals and draw a uniform random variable within its sub-interval. If the mode of the density lies in the interior of $(0, 1)$, then the grid method procedure works well.

The conditional posterior distribution of μ is

$$\pi(\mu|\sigma^2, \mathbf{c}, \mathbf{n}_1, \mathbf{n}_2) \propto \prod_{j=1}^K \left[\Phi(c_j) \Phi\left(\frac{c_j - \mu}{\sigma}\right) - \Phi(c_{j-1}) \Phi\left(\frac{c_{j-1} - \mu}{\sigma}\right) \right]^{n_{2j}} \frac{1}{1 + \mu^2},$$

and the parameter μ in $(-\infty, \infty)$ is transformed to the parameter ν in $(0, 1)$ using $\nu = e^\mu / (1 + e^\mu)$. The transformed conditional posterior distribution for ν is

$$\pi(\nu|\sigma^2, \mathbf{c}, \mathbf{n}_1, \mathbf{n}_2) \propto \prod_{j=1}^K \left[\Phi(c_j) \Phi\left(\frac{c_j - \text{logit}(\nu)}{\sigma}\right) - \Phi(c_{j-1}) \Phi\left(\frac{c_{j-1} - \text{logit}(\nu)}{\sigma}\right) \right]^{n_{2j}} \frac{1}{1 + \{\text{logit}(\nu)\}^2} \frac{1}{\nu(1-\nu)},$$

where $\text{logit}(\nu) = \log(\nu/(1-\nu))$ and we generate samples using the grid method.

The conditional posterior distribution of σ^2 is

$$\pi(\sigma^2|\mu, \mathbf{c}, \mathbf{n}_1, \mathbf{n}_2) \propto \prod_{j=1}^K \left[\Phi(c_j) \Phi\left(\frac{c_j - \mu}{\sigma}\right) - \Phi(c_{j-1}) \Phi\left(\frac{c_{j-1} - \mu}{\sigma}\right) \right]^{n_{2j}} \frac{1}{(1 + \sigma^2)^2},$$

and the parameter σ^2 is generated using the grid method on $(0, 1)$ after transformation to $\tau = \sigma^2 / (1 + \sigma^2)$, $0 < \tau < 1$. The transformed conditional posterior distribution for τ is

$$\pi(\tau|\mu, \mathbf{c}, \mathbf{n}_1, \mathbf{n}_2) \propto \prod_{j=1}^K \left[\Phi(c_j) \Phi\left(\frac{c_j - \mu}{\sqrt{\tau/(1-\tau)}}\right) - \Phi(c_{j-1}) \Phi\left(\frac{c_{j-1} - \mu}{\sqrt{\tau/(1-\tau)}}\right) \right]^{n_{2j}}.$$

Finally, the conditional posterior distributions of the boundaries c_j are given by

$$\begin{aligned} \pi(c_j|\mu, \sigma^2, \mathbf{n}_1, \mathbf{n}_2) \\ \propto \frac{e^{c_j}}{(1 + e^{c_j})^2} \left[\Phi(c_j) - \Phi(c_{j-1}) \right]^{n_{1j}} \left[\Phi(c_{j+1}) - \Phi(c_j) \right]^{n_{1,j+1}} \\ \times \left[\Phi(c_j) \Phi\left(\frac{c_j - \mu}{\sigma}\right) - \Phi(c_{j-1}) \Phi\left(\frac{c_{j-1} - \mu}{\sigma}\right) \right]^{n_{2j}} \left[\Phi(c_{j+1}) \Phi\left(\frac{c_{j+1} - \mu}{\sigma}\right) - \Phi(c_j) \Phi\left(\frac{c_j - \mu}{\sigma}\right) \right]^{n_{2,j+1}}, \end{aligned}$$

and we also transform the parameter c_j to parameter t_j in $(0, 1)$ using $t_j = e^{c_j} / (1 + e^{c_j})$, $j = 1, \dots, K-1$. We have $t_1, t_2, \dots, t_{K-1} \sim U(0, 1)$ such that $0 < t_1 < t_2 < \dots < t_{K-1} < 1$. Therefore, the transformed conditional posterior distribution for t_j is

$$\begin{aligned} \pi(t_j|\mu, \sigma^2, \mathbf{n}_1, \mathbf{n}_2) \propto \left[\Phi(\text{logit}(t_j)) - \Phi(\text{logit}(t_{j-1})) \right]^{n_{1j}} \left[\Phi(\text{logit}(t_{j+1})) - \Phi(\text{logit}(t_j)) \right]^{n_{1,j+1}} \\ \times \left[\Phi(\text{logit}(t_j)) \Phi\left(\frac{\text{logit}(t_j) - \mu}{\sigma}\right) - \Phi(\text{logit}(t_{j-1})) \Phi\left(\frac{\text{logit}(t_{j-1}) - \mu}{\sigma}\right) \right]^{n_{2j}} \\ \times \left[\Phi(\text{logit}(t_{j+1})) \Phi\left(\frac{\text{logit}(t_{j+1}) - \mu}{\sigma}\right) - \Phi(\text{logit}(t_j)) \Phi\left(\frac{\text{logit}(t_j) - \mu}{\sigma}\right) \right]^{n_{2,j+1}}, \end{aligned}$$

and the samples of t_j , $j = 1, \dots, K-1$ are drawn using the grid method on the conditions that $t_{j-1} < t_j < t_{j+1}$.

Table 1: Simulation results for AUCs using 100 simulated data with true AUC = 0.879

(1) Conventional binormal model								
Sample size	AUC	AB	RAB	RPMSE	C-CI	W-CI	C-HPD	W-HPD
(60, 30)	0.818	0.063	0.072	0.091	0.82	0.230	0.88	0.225
(45, 45)	0.820	0.060	0.068	0.082	0.87	0.200	0.90	0.196
(30, 60)	0.770	0.109	0.124	0.131	0.69	0.277	0.75	0.268
(600, 300)	0.863	0.017	0.019	0.024	0.84	0.061	0.86	0.060
(450, 450)	0.850	0.029	0.033	0.033	0.51	0.060	0.53	0.059
(1000, 100)	0.867	0.017	0.020	0.033	0.97	0.100	0.99	0.098
(10000, 100)	0.861	0.023	0.026	0.036	0.89	0.098	0.90	0.097
(15000, 100)	0.861	0.024	0.028	0.037	0.90	0.097	0.92	0.096
Average	0.839	0.043	0.049	0.058	0.811	0.140	0.841	0.137
(2) Binormal model under stochastic ordering								
Sample size	AUC	AB	RAB	RPMSE	C-CI	W-CI	C-HPD	W-HPD
(60,30)	0.830	0.052	0.059	0.074	0.82	0.186	0.830	0.183
(45,45)	0.827	0.053	0.060	0.073	0.88	0.180	0.890	0.177
(30,60)	0.800	0.079	0.090	0.097	0.71	0.209	0.780	0.206
(600,300)	0.871	0.011	0.013	0.019	0.96	0.054	0.980	0.054
(450,450)	0.862	0.017	0.019	0.022	0.85	0.053	0.870	0.052
(1000,100)	0.875	0.014	0.016	0.027	1.00	0.084	1.000	0.083
(10000,100)	0.873	0.017	0.019	0.028	0.94	0.079	0.940	0.079
(15000,100)	0.873	0.018	0.020	0.028	0.95	0.078	0.950	0.077
Average	0.851	0.033	0.037	0.046	0.889	0.115	0.905	0.114

AUC = area under the curve; AB = absolute bias; RAB = relative absolute bias; RPMSE = root posterior mean squared error; C-CI = coverage probability for 95% credible interval; W-CI = width of a credible interval; C-HPD = coverage probability HPD credible interval; W-HPD = width of the HPD credible interval; HPD = highest posterior density.

The convergence of the MCMC algorithm are assessed using the trace plots and autocorrelation plots. We use a single chain; therefore, convergence is checked using the Geweke test, which compares the mean of the initial 10% and the last 50% samples of the total iteration (Geweke, 1992).

The ROC curves are estimated by $\text{ROC}(u) = 1 - F_2(F_1^{-1}(1 - u))$, for $0 \leq u \leq 1$, and the AUC is obtained by $\text{AUC} = \int_0^1 \text{ROC}(u)du$ after estimating the parameters.

3. Simulation study

We conduct a simulation to compare the performance of the binormal model under stochastic ordering and the conventional binormal model. Metz and Pan (1999) discussed that the true population ROC curve does not have a hook, but the empirical ROC curve may have a hook in small sample size. The true ROC curve is a convex curve, but the estimated ROC curve from the simulated data with 5 categories (30 actually-negative and 40 actually-positive cases) has a hook in Figure 9 in Metz and Pan (1999). This population ROC curve has operating points (FPR, TPR) = (0.00025, 0.10), (0.000328, 0.25), (0.03040, 0.50), (0.28114, 0.85) (Dorfman and Berbaum, 1995) and the true AUC is 0.879 (Metz and Pan, 1999). The probabilities of 5 categories using the operating points are calculated as: $\mathbf{p}_1 = (0.71886, 0.25074, 0.030072, 0.000078, 0.00025)'$ for actually-negative cases and $\mathbf{p}_2 = (0.15, 0.35, 0.25, 0.15, 0.1)'$ for actually-positive cases.

We generate 100 simulated data from the multinomial distribution with probabilities $\mathbf{p}_1$ and $\mathbf{p}_2$ and sample sizes of (60, 30), (45, 45), (30, 60), (600, 300), (450, 450), (1000, 100), (10000, 100), (15000, 100). We then apply the binormal model under stochastic ordering and the conventional binormal model and estimate AUCs. The performance of the model are compared using the absolute

Figure 1: The estimated ROC curves (dash line) and 95% pointwise credible intervals (dotted line) derived from the conventional binormal model and the true ROC curve with $AUC = 0.879$ (solid line) from simulated data. ROC = receiver operating characteristic; AUC = curve and the area under the curve.

bias (AB), relative absolute bias (RAB) and root posterior mean squared error (RPMSE) for AUCs (Lee *et al.*, 2017). The RPMSE in r^{th} iteration is defined by $RPMSE_r = \sqrt{PSD_r^2 + AB_r^2}$, where PSD_r is the posterior standard deviation (PSD) in r^{th} simulated data. We also calculate the 95% credible interval (CI) and highest posterior density (HPD) credible interval for AUCs of each simulated data,

Figure 2: The estimated ROC curves (dash line) and 95% pointwise credible intervals (dotted line) derived from the binormal model under stochastic ordering and the true ROC curve with $AUC = 0.879$ (solid line) from simulated data. ROC = receiver operating characteristic; AUC = curve and the area under the curve.

and calculate the width of the credible interval and the coverage probability. Finally, we calculate the average of AB, RAB, and RPMSE, the width of credible interval and coverage probability to summarize the values for each simulated data (Hidirolou and You, 2016).

Simulation results of AUCs using 100 simulated data with true $AUC = 0.879$ are summarized in

Table 2: Artificial breast cancer data of film mammography screening

	1	2	3	4	5
Patients without breast cancer	11837	2839	1002	322	0
Patients with breast cancer	37	12	13	24	14

Table 3: Posterior mean, standard deviation, 95% credible intervals, and HPD CI of the area under the curve for the breast cancer data

	PM	PSD	95% CI	95% HPD CI
Binormal model under stochastic ordering	0.764	0.026	(0.714, 0.815)	(0.714, 0.814)
Conventional binormal model	0.695	0.044	(0.607, 0.777)	(0.612, 0.780)

PM = posterior mean; PSD = standard deviation; CI = 95% credible intervals; HPD = highest posterior density.

Table 1. The means of AUC in the binormal model under stochastic ordering are closer to 0.879 than those in the conventional binormal model. All measures for AB, RAB, RPMSE, the widths of the credible interval, and the widths of the HPD credible interval in each sample size are smaller in the binormal model under stochastic ordering than in the conventional binormal model. The coverage probabilities in the binormal model under stochastic ordering are closer to 0.95 than those in the conventional binormal model. Figure 1 and Figure 2 represent the estimated ROC curves derived from the binormal model under stochastic ordering and the conventional binormal model for 100 simulated data with the true ROC curve. The estimated ROC curves gets closer to the true ROC curve and the variation decreases as the sample size in both models increases. Many estimated ROC curves are improper ROC curves with a hook in the case of smaller number of samples in addition, a few estimated ROC curves have small degree of hook when the ratio of actually-negative to actually-positive cases is larger. Therefore, this simulation study indicates that the binormal model under stochastic ordering could estimate the proper ROC curve with a small bias despite sample sizes being small or significant variations in the sample size for actually-negative cases from actually-positive cases.

4. Real data application

Pisano *et al.* (2005) analyzed the diagnostic performance of film mammography for breast-cancer screening using the seven-point malignancy scale: 1 (definitely not malignant), 2 (almost definitely not malignant), 3 (probably not malignant), 4 (may be malignant), 5 (probably malignant), 6 (almost definitely malignant), 7 (definitely malignant). The estimated ROC curve using the conventional binormal model presented in Figure 1 in Pisano *et al.* (2005) and all ROC curves of film mammography represent improper ROC curves with a hook. Pesce *et al.* (2010) measured the FPF and TPF values from the ROC curve in panel D of Figure 1 which displays the obvious hook and simplified them to 5-category data. The probability of 5 categories were: $\mathbf{p}_1 = (0.74, 0.18, 0.06, 0.02, 0.00)'$ for actually-negative cases and $\mathbf{p}_2 = (0.43, 0.11, 0.09, 0.24, 0.13)'$ for actually-positive cases. Pesce *et al.* (2010) considered 100 actually-positive and 16,000 actually-negative cases. This is highly unbalanced and often occurs in the diagnostic test for a disease with a low prevalence. We create artificial data according to Pesce *et al.* (2010) (Table 2).

We apply the binormal model under stochastic ordering and the conventional binormal model. We draw 60,000 samples from the conditional posterior distribution for each parameter and select every 5th iterate after discarding 10,000 samples to obtain 10,000 samples for inference. All p-values of the Geweke test are greater than 0.1 and the effective sizes are higher than 9,000 in both models.

Table 3 summarizes the posterior mean (PM), PSD, 95% CI and HPD CI of the AUC for the

Figure 3: The ROC curves and the posterior densities of AUCs of the binormal model under stochastic ordering (dotted line) and the conventional binormal model (dash line) for the breast cancer data. ROC = receiver operating characteristic; AUC = curve and the area under the curve.

breast cancer data. The PM is 0.764, PSD is 0.026, 95% CI is (0.714, 0.815), and 95% HPD CI is (0.714, 0.814) in the binormal model under stochastic ordering. The PM is 0.695, PSD is 0.044, 95% CI is (0.607, 0.777), and 95% HPD CI is (0.612, 0.780) in the conventional binormal model. The PM from the binormal model under stochastic ordering is larger than the conventional binormal model because the fitted ROC curve obtained from the conventional binormal model has a hook. The PSD and the width of CI and HPD CI of the binormal model under stochastic ordering are smaller as compared to the conventional binormal model.

In Figure 3, we display the ROC curves and the posterior densities of AUCs of the binormal model under stochastic ordering and the conventional binormal model for breast cancer data. The circles represent empirical operating points. The fitted ROC curve obtained from the conventional binormal model are an improper ROC curve with a hook, but proper in the binormal model under stochastic ordering. The AUC distribution in the conventional binormal model highly vary and shift to the left of those in the binormal model under stochastic ordering.

5. Concluding remarks

In the case of low-prevalence diseases, the number of people who actually have a disease in screening diagnostic tests is generally lower compared to those without a disease. Pesce *et al.* (2010) mentioned that the probability of having a hook often increased with a small sample size, and the ratio of actually-negative to actually-positive cases is far from 1, based on experience. Therefore, it is important to assess the performance of the conventional binormal model and compare it with the model for estimating the proper ROC curve in sample size with large variations.

In this study, we apply the stochastic ordering method in a Bayesian hierarchical model to estimate the proper ROC curve and AUC when diagnostic test results are measured with discrete ordinal data and compare with the conventional binormal model and binormal model under stochastic ordering. The simulation study indicates that the binormal model under stochastic ordering can be used to estimate the proper ROC curve with a small bias despite the sample sizes being small or the large variation between actually-negative and actually-positive cases. In breast cancer data, the fitted ROC curve derived from the conventional binormal model represents an improper ROC curve with a hook, but proper in the binormal model under stochastic ordering. The distribution of AUC derived from the conventional binormal model varied greatly and shifted to the left in the binormal model under stochastic ordering. Therefore, it is appropriate to consider the binormal model under stochastic

ordering in the case of a sample size large variation between actually-negative and actually-positive groups.

Acknowledgements

This work was supported by a 2016 Research Funds of Andong National University grant.

References

- Bandos AI, Guo B, and Gur D (2017). Estimating the area under ROC curve when the fitted binormal curves demonstrate improper shape, *Academic Radiology*, **24**, 209–219.
- Bhattacharya B and Nandram B (1996). Bayesian inference for multinomial population under stochastic ordering, *Journal of Statistical Computation and Simulation*, **54**, 145–163.
- Dorfman DD and Alf 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 DD, Berbaum KS, Metz CE, Lenth RV, Hanley JA, and Abu Dagga H (1996). Proper receiver operating characteristic analysis: the bigamma model, *Academic Radiology*, **4**, 138–149.
- Gelfand AE and Kottas A (2001). Nonparametric Bayesian modeling for stochastic order, *Annals of the Institute of Statistical Mathematics*, **53**, 865–876.
- Gelman A (2006). Prior distribution for variance parameters in hierarchical models, *Bayesian Analysis*, **1**, 515–533.
- Gelman A, Jakulin A, Pittau MG, and Su YS (2008). A weakly informative default prior distribution for logistic and other regression models, *The Annals of Applied Statistics*, **2**, 1360–1383.
- Geweke J (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In Bernardo, JM, Berger, J, Dawid, AP and Smith, JFM (Eds.), *Bayesian Statistics* (Vol 4, pp. 169–194), Clarendon Press, Oxford.
- Gonçalves L, Subtil A, Oliveira MR, and Bermudez PDZ (2014). ROC curve estimation: An overview, *REVSTAT-Statistical Journal*, **12**, 1–20.
- Hanley JA (1988). The robustness of the “binormal” assumptions used in fitting ROC curves, *Medical Decision Making*, **8**, 197–203.
- Hanson TE, Kottas A, and Branscum AJ (2008). Modelling stochastic order in the analysis of receiver operating characteristic data: Bayesian non-parametric approaches, *Journal of Applied Statistics*, **57**, 207–225.
- Hidiroglou MA and You Y (2016). Comparison of unit level and area level small area estimators, *Survey Methodology*, **42**, 41–61.
- Hughes G and Bhattacharya B (2013). Symmetry properties of bi-normal and bi-gamma receiver operating characteristic curves are described by Kullback-Leibler divergences, *Entropy*, **15**, 1342–1356.
- Hwang BS and Chen Z (2015). An integrated Bayesian nonparametric approach for stochastic and variability orders in ROC curve estimation: an application to endometriosis diagnosis, *Journal of the American Statistical Association*, **110**, 923–934.
- Ishwaran H and Gatsonis AC (2000). A general class of hierarchical ordinal regression models with applications to correlated ROC analysis, *The Canadian Journal of Statistics*, **28**, 731–750.
- Johnson TD and Johnson VE (2006). A Bayesian hierarchical approach to multirater correlated ROC analysis, *Statistics in Medicine*, **25**, 1858–1871.
- Lee D, Nandram B, and Kim D (2017). Bayesian predictive inference of a proportion under a two-fold

- small area model with heterogeneous correlations, *Survey Methodology*, **17**, 69–92.
- Metz CE (1989). Some practical issues of experimental design and data analysis in radiological ROC studies, *Investigative Radiology*, **24**, 243–245.
- Metz CE, Herman BA, and Shen J (1998). Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data, *Statistics in Medicine*, **17**, 1033–1053.
- Metz CE and Pan X (1999). “Proper” binormal ROC curves: theory and maximum-likelihood estimation, *Journal of Mathematical Psychology*, **43**, 1–33.
- Molina I, Nandram B, and Rao JNK (2014). Small area estimation of general parameters with application to poverty indicators: a hierarchical Bayes approach, *The Annals of Applied Statistics*, **8**, 852–885.
- Mossman D and Peng H (2016). Using dual beta distributions to create “Proper” ROC curves based on rating category data, *Medical Decision Making*, **36**, 349–365.
- Nandram B, Bhatta D, Bhadra D, and Shen G (2013). Bayesian predictive inference of a finite population proportion under selection bias, *Statistical Methodology*, **11**, 1–21.
- Nandram B and Peiris TB (2018). Bayesian analysis of a ROC curve for categorical data using a skew-binormal model, *Statistics and Its Interface*, **11**, 369–384.
- Nandram B, Toto MCS, and Choi JW (2011). A Bayesian benchmarking of the Scott-Smith model for small areas, *Journal of Statistical Computation and Simulation*, **81**, 1593–1608.
- Nandram B and Yin J (2016). A nonparametric Bayesian prediction interval for a finite population mean, *Journal of Statistical Computation and Simulation*, **86**, 1–17.
- Peng F and Hall WJ (1996). Bayesian analysis of ROC curves using Markov-chain Monte Carlo methods, *Medical Decision Making*, **16**, 404–411.
- Pesce LL, Metz CE, and Berbaum KS (2010). On the convexity of ROC curves estimated from radiological test results, *Academic Radiology*, **17**, 960–968.
- Pisano ED, Hendrick E, Yaffe M, *et al.* (2005). Diagnostic performance of digital versus film mammography for breast-cancer screening, *The New England Journal of Medicine*, **353**, 1773–1783.
- Swets JA (1986). Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance, *Psychological Bulletin*, **99**, 181–198.
- Wang C, Turnbull BW, Gröhn YT, and Nielsen SS (2007). Nonparametric estimation of ROC curves based on Bayesian models when the true disease state is unknown, *Journal of Agricultural, Biological, and Environmental Statistics*, **12**, 128–146.