Review for Modelling stochastic order in the analysis of receiver operating characteristic data: Bayesian non-parametric approaches

Yu Zhu

June 10, 2022

1 Introduction

A ROC curve is a graphical measure for the accuracy of some continuous diagnostic test. And AUC is one of the popular information summarization methods of ROC curve. When AUC is 0.5, it indicates a random guess, or we define as the non-informative test. So a nature consequence of the order constraint appears that AUC should be greater than 0.5 to make the test meaningful. To model this stochastic order constraint, we can apply the parametric model to optimize the parameter under the constrain, or apply the non-parametric model. However, the distributions for two populations, for example, diseased and non-diseased (F0 and F1) often exhibit non-standard features such as multimodality and skewness that parametric models are not as flexible to capture. Non-parametric models can handle some other non-standard features that aren't known in advance. Bayesian inference also avoids constrained optimization since a prior restriction will imply the order in posteriori.Gelfand and Kottas (2001) proposed the general Bayesian non-parametric modeling under the stochastic order, and Hanson et al. (2008) proposed the methods of modeling stochastic order in the analysis of ROC under the Bayesian non-parametric framework. We reviewed the proposed two non-parametric models Dirichlet process mixtures (DPM) adn Mixtures of finite Polya tree (MPT) with stochastic order constraint, and performed some simulation study based on the DPM model.

2 ROC Curves

For some continuous test, we can define the true positive response probability (TPF) and the false positive response probability (FPF) under some cut-off value k as:

$$TPF(k) = Pr(Y \geq k|D = 1) = Pr(Y_1 \geq k) = 1 - F_1(k)$$
$$FPF(k) = Pr(Y \geq k|D = 0) = Pr(Y_0 \geq k) = 1 - F_0(k)$$

F_1 and F_0 denote the cumulative distribution function of test results in the diseased and non-diseased populations. ROC curve represents the plot $\{1 - F_0(k), 1 - F_1(k)\}$ for all cut-off values k. Denote $u = FPF(k) = 1 - F_0(k)$, we can present the ROC curve as:

$$ROC(u) = 1 - F_1(F_0^{-1}(1 - u)), u \in [0, 1]$$

ROC curve can measure the amount of separation between the distribution of test outcomes in the diseased and non-diseased populations. Under a non-informative test, the distributions of test results in two populations will be

completely overlapped and we will have $TPF(k) = FPF(k)$ for all k. The closer the ROC curve is to the point (0, 1), the more separated the distributions of test outcomes are, thus the better diagnostic accuracy we will obtain.

AUC is one of the information summarization for ROC curve. It represents the probability that the test outcome for a randomly chosen diseased subject exceeds the one exhibited by a randomly selected non-diseased individual. AUC = 0.5 suggests a non-informative test, and AUC = 1 indicates a perfect diagnostic test.

$$AUC = \int_0^1 ROC(u)du$$

A nature consequence of the stochastic order constraint is that $F_0 > F_1$ if and only if $ROC(u) > u$, which implies that $AUC > 0.5$. Hanson et al. proposed Bayesian nonparametric models taking this stochastic order constraint into consideration. The natural interpretation of stochastic order is that diseased people are more likely to test positive than non-diseased subjects for all k.

The distributions F_0 and F_1 often exhibit non-standard features such as mul-timodality and skewness, and parametric models are not as flexible to capture those features. So the motivation of applying the Bayesian non-parametric approaches is that non-parametric models can handle unspecified skewness and multimodality, or any other non-standard features that aren't known in advance. Two models are applied to incorporate the stochastic order constraint for F_0 and F_1: Dirichlet Process Mixtures (DPM) and Mixtures of Finite Polya Trees (MPT).

3 Dirichlet Process Mixtures

The support for F_0 and F_1 is taken to be the real line. Denote $D = \{\boldsymbol{x}_0, \boldsymbol{x}_1\}$ where $\boldsymbol{x}_0 = \{x_{01}, ..., x_{0n_0}\}$ is the data from non-diseased population and $\boldsymbol{x}_1 = \{x_{11}, ..., x_{1n_1}\}$ is the data from diseased population. Non-parametric prior probability models are proposed under the stochastic order restriction $F_1 \leq_{st} F_0$, i.e. $F_1(t) \leq F_0(t)$, for all $t \in R$.

$$F_l(t) \equiv F_l(t; H_l, \sigma^2) = \int N(t; \theta, \sigma^2) dH_l(\theta), l = 0, 1$$

Based on the fact that the $N(\theta, \sigma^2)$ distribution is stochastically ordered in θ for fixed σ^2, i.e., if $\theta_1 \leq \theta_2$, $N(\theta_1, \sigma^2) \leq_{st} N(\theta_2, \sigma^2)$. Thus, we would obtain the stochastic ordering for mixtures if the mixing distributions are stochastically ordered (Shaked and Shanthikumar (1994)).

$$H_0 \leq_{st} H_1$$
$$\text{then } F_0(t; H_0, \sigma^2) \leq_{st} F_1(t; H_1, \sigma^2)$$

To construct the stochastic order for the pair of mixing distributions (H_0, H_1), introduce the latent distribution functions H and G, and denote $H_0(t) = H(t)$ and $H_1(t) = H(t)G(t)$. Now the stochastically ordered DPM models are defined as below:

$$F_0(t; H, \sigma^2) = \int N(t; \theta, \sigma^2) dH(\theta)$$
$$F_1(t; H, G, \sigma^2) = \int N(t; max(\theta, \phi), \sigma^2) dH(\theta) dG(\phi)$$
$$H \sim DP(\alpha_H, N(\mu_H, \tau_H^2))$$
$$G \sim DP(\alpha_G, N(\mu_G, \tau_G^2))$$

The hyper-parameters are defined as $\psi = (\alpha_H, \mu_H, \tau_H^2, \alpha_G, \mu_G, \tau_G^2)$

4 A Mixtures of Polya Tree

Define H_0 and H_1 under the same structure as above and introduce the latent distribution functions H and G. We directly model F_0 and F_1 as $F_0 = H_0$ and $F_1 = H_1$.

$$F_0(t) = H_0(t) = H(t)$$
$$F_1(t) = H_1(t) = H(t)G(t)$$

Now the mixture of finite PT priors are assigned:

$$H \sim \int FPT(c_H, H_{\boldsymbol{\theta}_H}) dP_H(c_H, \boldsymbol{\theta}_H)$$
$$G \sim \int FPT(c_G, G_{\boldsymbol{\theta}_G}) dP_G(c_G, \boldsymbol{\theta}_G)$$

H is centered randomly at $H_{\boldsymbol{\theta}_H} = N(\mu_H, \tau_H^2)$ with $\boldsymbol{\theta}_H = (\mu_H, \tau_H^2)$ and G is centered randomly at $G_{\boldsymbol{\theta}_G} = N(\mu_G, \tau_G^2)$ with $\boldsymbol{\theta}_G = (\mu_G, \tau_G^2)$. The levels of the finite PTs defining H and G are set to J_H and J_G, and we fix $J_H = J_G \equiv J$. Bias-Variance tradeoff exists in deciding the value of J: increasing J to J +1 essentially doubles the number of conditional probabilities and decrease the bias, while also increases overall variability and can reduce the predictive ability.

5 Simulation Study

We randomly select $n_0 = n_1 = 300$ observations for each population, where

$$x_{0i} \sim 0.5N(0,1) + 0.1N(1,1) + 0.4N(-5,1), j = 1, 2, ..., n_1$$
$$x_{1j} \sim 0.5N(0,4) + 0.5N(1,1), i = 1, 2, ..., n_0$$

Model can be expressed in hierarchical form with latent mixing parameters $\boldsymbol{\theta} = \{\theta_i : i = 1, ..., n_0, n_0+1, ..., n_0+n_1\}$ and $\boldsymbol{\phi} = \{\phi_i : i = 1, ..., n_1\}$ as below:

$$x_{0i}|\theta_i, \sigma^2 \overset{\text{ind}}{\sim} N(\theta_i, \sigma^2), i = 1, 2, ..., n_0$$
$$x_{1j}|\theta_{n_0+j}, \phi_j, \sigma^2 \overset{\text{ind}}{\sim} N(max(\theta_{n_0+j}, \phi_j), \sigma^2), j = 1, 2, ..., n_1$$
$$\theta_i|H \overset{\text{IID}}{\sim} H, i = 1, ..., n_0, n_0+1, ..., n_0+n_1$$
$$\theta_j|G \overset{\text{IID}}{\sim} G, j = 1, 2, ..., n_1$$
$$H, G|\alpha_H, \mu_H, \tau_H^2, \alpha_G, \mu_G, \tau_G^2 \sim DP(\alpha_H, N(\mu_H, \tau_H^2))DP(\alpha_G, N(\mu_G, \tau_G^2))$$

The introduction of the additional mixing parameters θ_{n_0+j}, allows us to retain the first-stage conditionally independent specification in the hierarchical model after marginalizing in model the random distribution functions H and G over their DP priors.

The priors that I set up are:

$$\mu_H \sim N(-2,4), \mu_G \sim N(0.5,4)$$
$$\tau_H^2 \sim IG(3,10), \tau_G^2 \sim IG(3,4)$$
$$\sigma^2 \sim IG(3,4)$$
$$\alpha_H, \alpha_G \sim G(2,0.9)$$

Let $n_\theta^*(\leq n_0+n_1)$ with $\{\theta_l^* : l=1,...,n_\theta^*\}$ and $n_\phi^*(\leq n_1)$ with $\{\phi_l^* : l=1,...,n_\phi^*\}$ be the number of and values of the distinct components in $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ respectively, If $\alpha_H \sim G(a_{\alpha.H}, b_{\alpha.H})$ and $\alpha_G \sim G(a_{\alpha.G}, b_{\alpha.G})$, we can approximate the expected values for n_θ^* and n_ϕ^* as:

$$E(n_\theta^*) \approx a_{\alpha.H} b_{\alpha.H}^{-1} log(1+(n_0+n_1)a_{\alpha.H}^{-1} b_{\alpha.H}) \approx 13$$
$$E(n_\phi^*) \approx a_{\alpha.G} b_{\alpha.G}^{-1} log(1+n_1 a_{\alpha.G}^{-1} b_{\alpha.G}) \approx 12$$

The posterior simulation can be performed from $p(\boldsymbol{\theta}, \boldsymbol{\phi}, \sigma^2, \boldsymbol{\psi}|D)$ by integrating H and G over their DP priors:

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}, \sigma^2, \boldsymbol{\psi}|D) \propto$$
$$\prod_{i=1}^{n_0} N(x_{0i}; \theta_i, \sigma^2) \prod_{j=1}^{n_1} N(x_{1j}; max(\theta_{n_0+j}, \phi_j), \sigma^2) p(\sigma^2) p(\boldsymbol{\theta}|\alpha_H, \mu_H, \tau_H^2) p(\boldsymbol{\phi}|\alpha_G, \mu_G, \tau_G^2) p(\boldsymbol{\psi})$$

The posterior sampling algorithm is similar to what we learned from the lecture, except for sampling $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. We need extra Metropolis step to sample the pair of (θ_{n_0+j}, ϕ_j). The algorithm is as below:

1. Sample $\theta_i|\{\theta_l : l \neq i\}, \alpha_H, \mu_H, \tau_H^2, D$ for $i : 1,...,n_0$ with

$$p(\theta_i|\{\theta_l : l \neq i\}, \alpha_H, \mu_H, \tau_H^2, D) = \frac{\alpha_H q_0^\theta}{\alpha_H q_0^\theta + \sum_{j=1}^{n_0^{*-}} n_{0j}^- q_j^\theta} h(\theta_i|\alpha_H, \mu_H, \tau_H^2, \sigma^2, x_{0i}) + \sum_{j=1}^{n_0^{*-}} \frac{n_{0j}^- q_j}{\alpha_H q_0^\theta + n_{0j}^- q_j^\theta} \delta_{\theta_j^{*-}}(\theta_i)$$

2. Sample $(\theta_{n_0+j}, \phi_j)|\{\theta_{n_0+l} : l \neq j\}, \{\phi_l : l \neq j\}, \alpha_G, \mu_G, \tau_G^2, \alpha_H, \mu_H, \tau_H^2, D$ for $j : 1,...,n_1$ with Metropolis step that

$$p(\theta_{n_0+j}|\{\theta_{n_0+l} : l \neq j\}, \alpha_H, \mu_H, \tau_G^2, D) = \frac{\alpha_H q_0^\theta}{\alpha_H q_0^\theta + \sum_{j=1}^{n_1^{*-}} n_{1j}^- q_j^\theta} h(\theta_{n_0+j}|\alpha_H, \mu_G, \tau_G^2, \sigma^2, x_{1i}) + \sum_{j=1}^{n_1^{*-}} \frac{n_{1j}^- q_j}{\alpha_H q_0^\theta + n_{1j}^- q_j^\theta} \delta_{\theta^{*-}}(\theta_n$$

$$p(\phi_j|\{\phi_l : l \neq j\}, \alpha_G, \mu_G, \tau_G^2, D) = \frac{\alpha_G q_0^\phi}{\alpha_G q_0^\phi + \sum_{j=1}^{n_1^{*-}} n_{1j}^- q_j^\phi} h(\phi_j|\alpha_G, \mu_G, \tau_G^2, \sigma^2, x_{1i}) + \sum_{j=1}^{n_1^{*-}} \frac{n_{1j}^- q_j}{\alpha_G q_0^\phi + n_{1j}^- q_j^\phi} \delta_{\phi^{*-}}(\phi_j)$$

   Accept the pair (θ_{n_0+j}, ϕ_j) with probability $min\{1, \frac{N(x_{1j}; max(\theta_{n_0+j}^{new}, \phi_j^{new}), \sigma^2)}{N(x_{1j}; max(\theta_{n_0+j}^{old}, \phi_j^{old}), \sigma^2)}\}$

3. Sample the rests following the structure in the slides

We not only interested in the posterior point estimates for the mixture distribution, but more importantly, the full posterior inference for $F_0(t; H, \sigma^2)$ and $F_1(t; H, G, \sigma^2)$. We require the posteriors of H and G. We can sample from these posteriors obtained by augmenting the MCMC algorithm discussed in the lecture. Specifically, based on Antoniak (1974),

$$p(H, G, \boldsymbol{\theta}, \boldsymbol{\phi}, \sigma^2, \boldsymbol{\psi}) = p(H|\boldsymbol{\theta}, \alpha_H, \mu_H, \tau_H^2) p(G|\boldsymbol{\phi}, \alpha_G, \mu_G, \tau_G^2) p(\boldsymbol{\theta}, \boldsymbol{\phi}, \sigma^2, \boldsymbol{\psi}, D)$$

where $p(H|\boldsymbol{\theta}, \alpha_H, \mu_H, \tau_H^2)$ denotes a DP distribution with precision parameter $\alpha_H + n_0 + n_1$ and base distribution; $p(G|\boldsymbol{\phi}, \alpha_G, \mu_G, \tau_G^2)$ denotes a DP distribution with precision parameter $\alpha_G + n_1$ and base distribution. As for the ROC curve, the inference of ROC cure is based on a grid of $u \in (0, 1)$. Invert $F_{ob}(t_m) : m = 1, ..., M$ each posterior realization from the random distribution function $F_0(.; H, \sigma^2)$ to obtain sample $\eta_b(u) : b = 1, ..., B$ from the posterior $F_0^{-1}(1 - u; H, \sigma^2)$. The we can compute

$$1 - F_1(\eta_b(u); H_b, G_b, \sigma_b^2) = 1 - \int\int N(\eta_b(u); max(\theta, \phi), \sigma_b^2) dH_b(\theta) dG_b(\phi)$$
$$ROC(; H, G, \sigma^2) = 1 - F_1(F_0^{-1}(1 - u; H, \sigma^2), H, G, \sigma^2)$$

Based on 5000 Burn-ins and 5000 iterations, we obtained the posterior predictive distribution and the inference for ROC curve with 95% interval estimates.

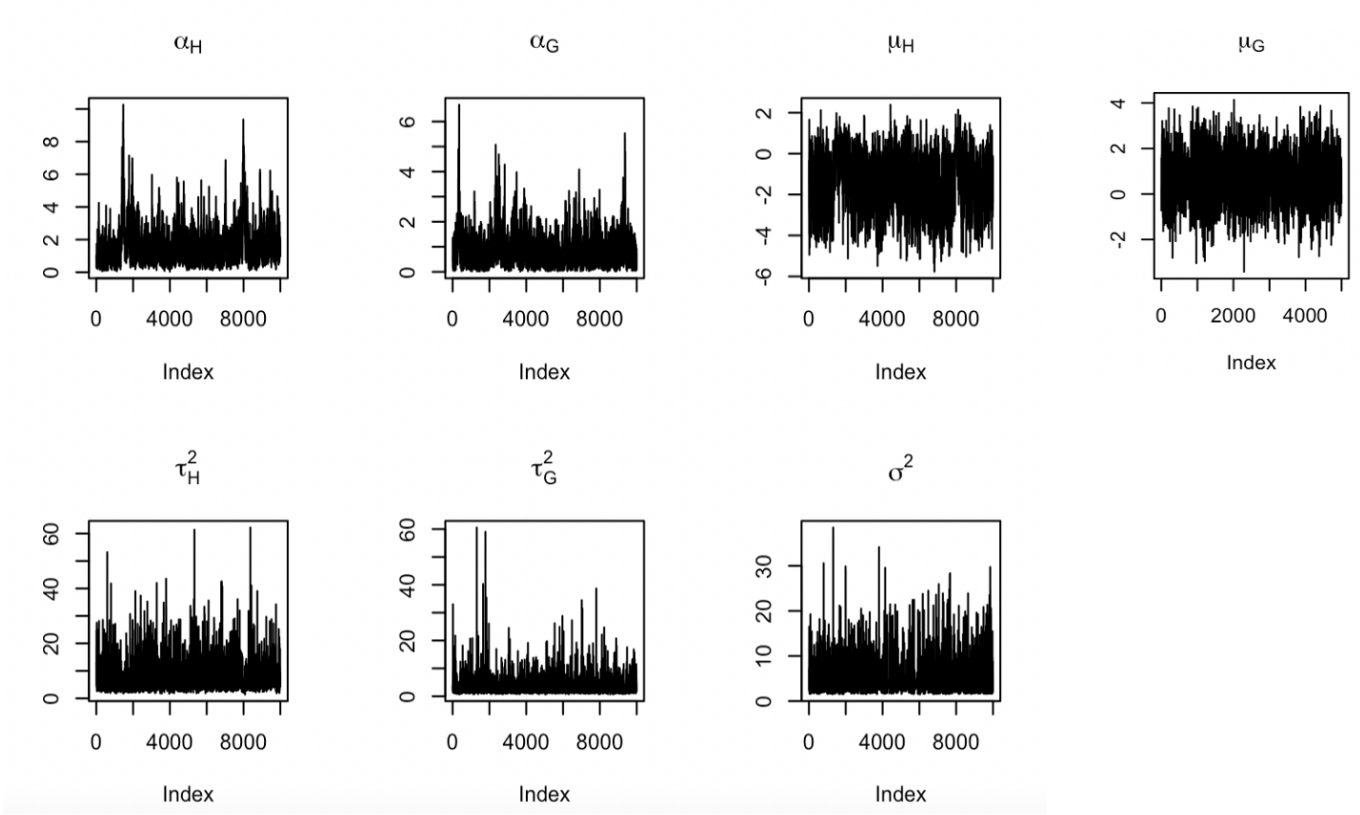

Figure 1: Traceplots

We plot the posterior predictive density for two distributions, and we can observe that the DPM model fit the data, catching the modes but does not capture the abrupt change pretty well. Comparing with the case study in the paper, it matches the results that comparing with DPM, MPT is more jagged and can catch the abrupt change in the data. DPM is smoother comparing with MPT expecially when the priors on c_H and c_G are large. The point estimate for AUC is 0.765 with interval estimate of (0.729, 0.799), and it's obviously larger than 0.5.

The Figure 3 shows the ROC curve under the true data in black, and it's almost overlapped with the lower bound of the interval. Based on the true data, AUC is 0.731. It's smaller than the posterior estimated AUC and bounded in that interval. The DPM model provides the better AUC in this case.

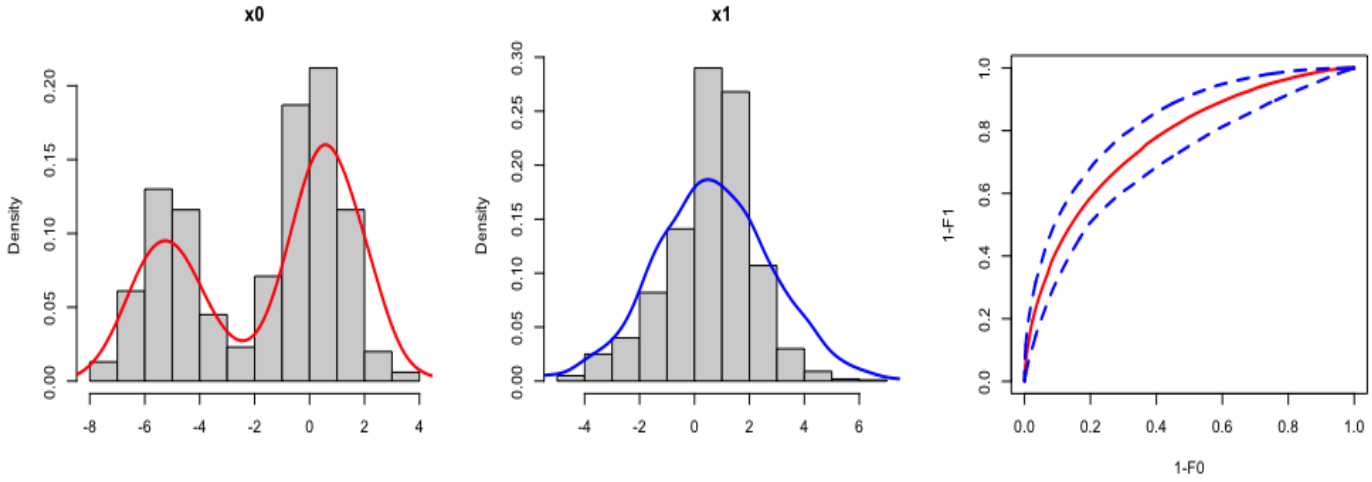

Figure 2: Posterior Predictive and ROC

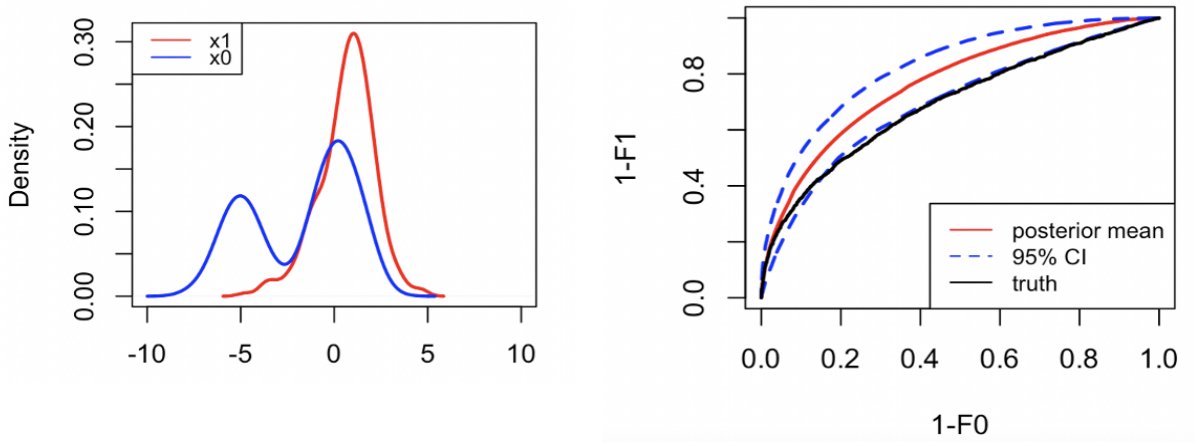

Figure 3: The ROC Curve Comparisons

6 Conclusion

We implemented the DPM model with stochasic order constraint based on the simulation study under two mixture of Normals. The means are generated in close values to make two distributions hard to separate. The model works well under the informative prior and fit the data well except for barely capturing the abrupt changes in the data. And the inference for ROC curve also shows that DPM model did a good job since now the test becomes even more informative comparing with the truth.

7 Future Work

For future work, I want to explore and compare the models' performance when sample size is small. When I used sample size equal to 500 in this simulation study, I found that the prior setting is quite influential to the posterior inference during the sensitivity analysis. It might due to the Metropolis step in the simulation algorithm. So the situation can be even worse when sample size is small. The MPT model may not be that sensitive to the sample size, but more simulation studies are wait to be explored.

Furthermore, the real case can include the covariates such as age and gender, and we can explore DDP with stochastic order constraint. de Carvalho et al. (2013) proposed DDP models on ROC regression, while they did not take the stochastic order constraint into consideration.

References

de Carvalho, V. I., Jara, A., Hanson, T. E., and de Carvalho, M. (2013). Bayesian nonparametric roc regression modeling. *Bayesian Analysis*, 8:623–646.

Gelfand, A. and Kottas, A. (2001). Nonparametric Bayesian Modeling for Stochastic Order. *Annals of the Institute of Statistical Mathematics*, 53(4):865–876.

Hanson, T. E., Kottas, A., and Branscum, A. J. (2008). Modelling stochastic order in the analysis of receiver operating characteristic data: Bayesian non-parametric approaches. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(2):207–225.