Bayesian semi-parametric ROC analysis

Alaatin Erkanli1,*,†, Minje Sung2,†, E. Jane Costello3 and Adrian Angold3

1Department of Biostatistics and Bioinformatics, Duke University Medical School, Box 3454, Durham, NC 27710, U.S.A.
2Department of Business Administration, Ajou University, Suwon, Kyungki-Do, Korea
3Department of Psychiatry and Behavioral Sciences, Duke University Medical School, Developmental Epidemiology Center, Box 3454, Durham, NC 27710, U.S.A.

SUMMARY

This paper describes a semi-parametric Bayesian approach for estimating receiver operating characteristic (ROC) curves based on mixtures of Dirichlet process priors (MDP). We address difficulties in modelling the underlying distribution of screening scores due to non-normality that may lead to incorrect choices of diagnostic cut-offs and unreliable estimates of prevalence of the disease. MDP is a robust tool for modelling non-standard diagnostic distributions associated with imperfect classification of an underlying diseased population, for example, when a diagnostic test is not a gold standard. For posterior computations, we propose an efficient Gibbs sampling framework based on a finite-dimensional approximation to MDP. We show, using both simulated and real data sets, that MDP modelling for ROC curve estimation closely parallels the frequentist kernel density estimation (KDE) approach. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: ROC curve; diagnostic test; mixtures of Dirichlet process priors; semi-parametric Bayesian methods; Gibbs sampling; kernel density estimation

1. INTRODUCTION

In medical testing, the results of a screening test usually do not have a Gaussian or a symmetric distribution. For example, depression scale scores in an adolescent population [1] are known to follow a J-shaped distribution, having a high probability concentration around lower values (typically zero). The standard approach to receiver operating characteristic (ROC) estimation is the binormal method [2], a parametric approximation of the ROC curve based on the assumption of univariate normality for a test score \( X \) in the diseased and non-diseased

*Correspondence to: A. Erkanli, Department of Biostatistics and Bioinformatics, Duke University Medical School, Box 3454, Durham, NC 27710, U.S.A.
†E-mail: al@psych.duhs.duke.edu
‡E-mail: sungmj@ajou.ac.kr

Contract/grant sponsor: NIH, U.S.A.; contract/grant number: MH-66497

Copyright © 2006 John Wiley & Sons, Ltd.

Received 2 June 2004
Accepted 30 November 2005
populations [3]. The program LABROC4 can be used to fit smooth ROC curves [4], however, the form of resulting ROC curves is always binormal. More recently, Sorribas et al. [5] introduced an ROC analysis based on the S-distribution family [6], which is described in terms of a non-linear differential equation solution with appropriate boundary conditions dependent on the sample moments. However, it requires potentially time consuming numerical integration. Like the binormal ROC, it also assumes a unimodal score distribution, an assumption that may be violated by the actual data.

While a non-parametric ROC estimate can easily be obtained using the empirical distributions of the samples obtained from the diseased and non-diseased populations, the associated ROC curve would be too primitive and difficult to interpret [7]. An alternative approach is to use kernel density estimation (KDE) which yields a smooth ROC curve [8, 9]. In this paper, we develop a semi-parametric Bayesian approach to estimating ROC curves [10] using mixtures of Dirichlet process priors (MDP) [11–14]. Our approach parallels very closely KDE, although there are some fundamental differences between the two. In the KDE approach, one has to choose the bandwidth (among other things) to strike a balance between smoothness and model-fit. In the Bayesian approach one has to specify appropriate prior distributions for the smoothness parameter(s) associated with the MDP [15] and parameters associated with the component mixtures. The precision parameters associated with each component of the mixture in the MDP framework provide generalizations over the single bandwidth parameter used in the KDE approach, allowing more flexibility in approximating the local densities more closely. While the accuracy of the KDE approach relies upon asymptotics which needs large samples, the accuracy of the Bayesian semi-parametric approach is assessed a posteriori by combining a high-dimensional likelihood function with the prior distribution of the model and tuning parameters irrespective of the sample size. Prior distributions can take into account subjective beliefs about the accuracy of a screen test based on expert opinions or historical information. In contrast, the KDE approach cannot utilize any form of prior information. It is, however, relatively easy to obtain non-parametric estimates of ROC curves utilizing the KDE approach. In contrast, the MDP approach depends on intensive computations and can be time consuming. The advantage, however, is that the posterior uncertainty about the ROC curve can be assessed probabilistically using the posterior predictive distribution of the ROC curve, whereas the KDE approach relies upon large sample frequentist coverage (or confidence intervals) which do not have a probabilistic interpretation. Moreover, Markov Chain Monte Carlo (MCMC) methods, in particular the Gibbs sampler [16], are very efficient tools that can aid in posterior computations and can be implemented in freely available WinBUGS [17] software. In Section 3, we compared the Bayesian and KDE approaches. The details of our modelling approach are described in Section 2; its accuracy is illustrated in Section 3 using simulated and real data sets; and conclusions and future proposal are discussed in Section 4. Technical developments of the MDP approach leading to semi-parametric Bayesian estimation of ROC curves are provided in Appendix A. The details of the Gibbs sampler are described in Appendix B, and WinBUGS [17] codes for fitting these models are given in Appendix C.

2. A SEMI-PARAMETRIC MODEL FOR ROC ANALYSIS

A generalization of the binormal [2–4] model for ROC curve estimation is achieved by relaxing the parametric assumptions about the underlying distribution of test scores conditional upon
disease status. A fully parametric model may not be adequate to capture the changes in the disease (signal) and non-disease (noise+signal) processes. As pointed out by Gelfand [18], in cases where parametric assumptions are too restrictive, a semi-parametric model can be developed by utilizing non-parametric specifications of some components of the model. The application of MDP for this purpose has been demonstrated in the context of density estimation [12], ordinal data [19], multivariate regression [14], survival analysis [20], and linear regression [21–23]. Recently, Ishwaran and James [24] described a computationally attractive approximation to the Dirichlet process that can be implemented in WinBUGS [17]. Building on these developments, we model the outcomes (e.g. test scores obtained from a screening device) in the diseased and non-diseased populations using MDP [13] methodology.

Briefly, the application of MDP to random samples from the diseased \( D = 1 \) and non-diseased \( D = 0 \) populations yields posterior predictive densities \( f_D(x \mid X_D) \) and predictive distribution functions \( F_D(x \mid X_D) \), where \( X_D \) is the vector of the observed screen scores in the diagnostic population \( D \). The elements of \( X_D \) are assumed to follow the normal hierarchical MDP model described in Appendix A (equation (A1)), and as a result of the Bayesian computations facilitated by the MCMC integration (Gibbs sampling), the posterior predictive densities \( f_D(x \mid X_D) \) (and the distribution functions \( F_D(x \mid X_D) \)) are approximated by the finite-mixtures as shown in equation (A4). We let \( Y \) denote a binary decision variable for a randomly selected subject that takes a value of 1 if the subject's test score exceeds a cut-off, and 0 otherwise. Then the posterior predictive true positive rate (TPR or sensitivity) and false positive rate (FPR or 1-specificity) are defined as

\[
\lambda(c \mid X_i) = \Pr(Y = 1 \mid X_i) = \Pr(X \geq c \mid X_i) = 1 - F_i(c \mid X_i)
\]

and

\[
\gamma(c \mid X_0) = \Pr(Y = 1 \mid X_0) = \Pr(X \geq c \mid X_0) = 1 - F_0(c \mid X_0)
\]

where

\[
F_D(c \mid X_D) = \int_{-\infty}^{c} f_D(x \mid X_D) \, dx
\]

The predictive ROC curve is obtained by varying the cut-off \( c \) and plotting \( \lambda \) against \( \gamma \). Because each \( F_D(x \mid X_D) \) for \( D = 0 \) and \( D = 1 \) has a finite mixture form, the predictive TPR and FPR in (1) and (2) are also described as finite-mixtures of component TPRs and FPRs, as illustrated in Appendix A. By solving (1) and (2) for the cut-off \( c \) an equivalent representation of the predictive ROC curve can be obtained in terms of quantiles

\[
F^{-1}_1(q_1 \mid X_i) = F^{-1}_0(q_0 \mid X_0)
\]

Under the MDP model we can also compute the area under the ROC curve (AUC) as

\[
\text{AUC} = \int [1 - \gamma(c \mid X_0)] \, d\lambda(c \mid X_i) = \int_{-\infty}^{\infty} \left[ \int_{-\infty}^{c} dF_0(x \mid X_0) \right] \, dF_i(x \mid X_i)
\]

which also has a finite mixture form (with double mixing weights) comprised of component AUCs (see equation (A12) in Appendix A) related to those obtained under binormal.
theory. The posterior uncertainties about TPR, FPR, ROC curve, and AUC are assessed using prediction intervals obtained from the posterior simulations based on the Gibbs sampling approach described in Appendix B.

3. APPLICATIONS

3.1. Example 1

To illustrate the accuracy of the MDP approach, the scores $X$ of an imaginary screening test were simulated from two diagnostic populations (high; $D=1$, and low; $D=0$):

1) $n=1600$ subjects in the high group were simulated from the mixture of three normal distributions

$$X_{t_D=1} \sim 0.125N(45,25) + 0.25N(65,25) + 0.625N(85,49)$$

2) $n=1500$ subjects in the low group were simulated from a mixture of four normal distributions

$$X_{t_D=0} \sim 0.13N(80,25) + 0.13N(65,49) + 0.34N(50,25) + 0.4N(35,16)$$

As shown in Appendix A, we chose a conjugate normal-gamma baseline prior for the components of model parameters $\{\theta_k\}$. To ensure identifiability, we used non-informative but proper, and ordered, priors for $\{\mu_k\}$ such that $\mu_1 < \mu_2 < \cdots < \mu_C$, described by independently and normally distributed increments $\delta_k = \mu_k - \mu_{k-1} \sim N(0,100)$. The precision parameters $\{\tau_k\}$ were assumed to be independently and identically distributed as Gamma $(0.01, 0.1)$, where $\sigma_k = 1/\sqrt{\tau_k}$.

In the MDP model the precision parameter $\alpha$ (the concentration parameter of the Dirichlet process) strongly influences the number of mixture components $C$. Technical details about the choice of priors for $\alpha$ and the choice of $C$ are in provided in Appendix B. They are based on Taylor series approximations for the marginal prior mean and prior variance of the number of distinct clusters $C \leq C^*$ (see Reference [25] for details, also see Escobar and West [12] for an alternative approach). A common choice for the prior of $\alpha$ is a Gamma $(a_0, b_0)$ distribution with $a_0 > 0$ and $b_0 > 0$. In general, prior distributions that support large values of $\alpha$ imply strong prior belief in the baseline measure $G_0$ and induce a smaller number of component mixtures concentrating around $G_0$. Conversely, prior distributions that support smaller values of $\alpha$ represent weak prior beliefs about $G_0$ and induce a large number of component mixtures. A prior distribution that is improper or too spread out over $(0, \infty)$ is very likely to create convergence problems in the MCMC sampler. Using equations (B4) and (B5) in Appendix A, we computed the values of the hyperparameters $a_0$ and $b_0$ to have a prior mean of 5, and prior variance of 10, for the number of distinct clusters $C^*$ (see Figure 3). We utilized the SOLVER function in Microsoft Excel [26] to find the optimal values for $a_0$ and $b_0$ (3.37, 5.28) for the high group and (3.36, 5.21) for the low group, respectively.

We also specified an upper bound for the number of mixtures as $C=10$, which we determined to be sufficient to pick up the variation patterns in the simulated data. The computations were implemented in the WinBUGS programming environment (code provided in Appendix C). We monitored five parallel chains using 5000 iterations for burn-in
and 10,000 additional iterations for inferences. Convergence was assessed visually by monitoring the dynamic traces of Gibbs iterations and by computing the Gelman–Rubin [27] convergence statistic. Computations were implemented in a Dell PC with dual 2.4 Xeon Pentium 4 processors and 1 GB memory. The posterior predictive samples of X were generated from the posterior mixtures using the last 1000 simulated values of the parameters \( \{ \mu_k \}, \{ \sigma_k^2 \} \) and associated weights \( \{ w_k \} \). S-Plus version 6.2 [28] was used to generate the posterior predictive samples and graphics.

Figure 1 shows the predictive density estimates for each group that were obtained using (1) the MDP approach, (2) the binormal approach, and (3) the KDE approach, plus histograms of the simulated data. For KDE, we used the optimal bandwidth [4, 13] given by

\[
\text{Bandwidth} \approx 0.9 \min(\text{SD, IQR}/1.34)n^{-1/3} 
\]

where \( n \) represents the sample size, SD is the standard deviation, and IQR is the inter-quartile range. Both MDP and KDE closely approximated the underlying mixture structure. In fact, the estimated ROC curve obtained using MDP is closer to that obtained using KDE. Table I

![Figure 1. Estimates of predictive densities of X based on MCMC simulations: MDP, Normal models, and KDE. Actual data are presented by the histograms.](image)
Table I. Posterior means of the mixing weights and parameters of the mixture (where $\sigma_k = 1/\sqrt{\nu_k}$).

<table>
<thead>
<tr>
<th>$k$</th>
<th>$\omega_k$</th>
<th>$\mu_k$</th>
<th>$\sigma_k$</th>
<th>$\omega_k$</th>
<th>$\mu_k$</th>
<th>$\sigma_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.000</td>
<td>0.37</td>
<td>1.00</td>
<td>0.000</td>
<td>0.56</td>
<td>1.01</td>
</tr>
<tr>
<td>2</td>
<td>0.002</td>
<td>34.77</td>
<td>1.07</td>
<td>0.389</td>
<td>33.67</td>
<td>1.54</td>
</tr>
<tr>
<td>3</td>
<td>0.118</td>
<td>44.44</td>
<td>4.64</td>
<td>0.344</td>
<td>35.85</td>
<td>2.48</td>
</tr>
<tr>
<td>4</td>
<td>0.034</td>
<td>57.25</td>
<td>1.39</td>
<td>0.096</td>
<td>48.20</td>
<td>2.92</td>
</tr>
<tr>
<td>5</td>
<td>0.147</td>
<td>63.58</td>
<td>3.07</td>
<td>0.093</td>
<td>52.32</td>
<td>1.29</td>
</tr>
<tr>
<td>6</td>
<td>0.127</td>
<td>72.00</td>
<td>2.15</td>
<td>0.061</td>
<td>56.02</td>
<td>7.62</td>
</tr>
<tr>
<td>7</td>
<td>0.103</td>
<td>77.39</td>
<td>1.40</td>
<td>0.012</td>
<td>67.36</td>
<td>1.57</td>
</tr>
<tr>
<td>8</td>
<td>0.130</td>
<td>83.89</td>
<td>2.01</td>
<td>0.003</td>
<td>79.02</td>
<td>2.75</td>
</tr>
<tr>
<td>9</td>
<td>0.186</td>
<td>87.98</td>
<td>1.72</td>
<td>0.001</td>
<td>85.79</td>
<td>1.14</td>
</tr>
<tr>
<td>10</td>
<td>0.151</td>
<td>93.00</td>
<td>1.75</td>
<td>0.000</td>
<td>93.17</td>
<td>1.10</td>
</tr>
</tbody>
</table>

![Graph](image)

Figure 2. Estimates of TPR, FPR and ROC based on MCMC simulations: MDP, Normal models, and KDE.

shows the posterior means of the mixing weights and parameters in the high and low diagnostic groups. In the high group, about 96.4 per cent were drawn from seven distinctive mixture components supporting 7–8 clusters. In the low group, there were two dominating clusters with the associated weights 38.9 and 34.4 per cent, and three additional clusters with weights higher than 5 per cent. Our semi-parametric Bayesian model worked equally well in the low (non-diseased) group.

The posterior predictive estimates of FPR, TPR, and ROC under the binormal, KDE, and MDP methods (respectively) are compared in Figure 2. Note that the shape of the ROC curves based on the MDP (—) is quite different from the ROC curve obtained using the
binormal method (-- --), but is very similar to the KDE estimate (---). For some subregions of TPR and FPR, the 95 per cent credible interval (dotted line) around the MDP estimate does not even contain the mean ROC curve based on the binormal method. If these regions were of immediate interest to a diagnostician, the inferences based on a binormal method might lead to an incorrect cut-off value and erroneous estimates of the sensitivities and specificities.

The fundamental differences between the MDP and binormal models are also apparent in posterior AUC estimates. The posterior mean AUC for the binormal model was 0.871, but the posterior mean for the MDP model was 0.853 with a 95 per cent credible interval of (0.841, 0.865) which does not even contain the posterior (mean) estimate for the binormal model. Thus, in this example, we might conclude that the binormal model overestimates the AUC, but the MDP approach provides a more accurate estimate of the AUC.

Figure 3 compares the prior and posterior distributions of \( \alpha \), and the number of distinct clusters \( C^* \) using the priors \( \alpha \sim \text{Gamma}(3.37, 5.28) \) for the high group, and \( \alpha \sim \text{Gamma}(3.66, 5.21) \) for the low group, respectively. Figure 3 shows that the MDP model supports 8 or 9 mixture components for the high group, and 5 or 6 components for the low group.
Table II. Posterior means of the mixing weights and parameters of the mixture.

<table>
<thead>
<tr>
<th></th>
<th>High group</th>
<th></th>
<th>Low group</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$w_k$</td>
<td>$\mu_k$</td>
<td>$\sigma_k$</td>
<td>$w_k$</td>
</tr>
<tr>
<td>1</td>
<td>0.000</td>
<td>0.29</td>
<td>1.00</td>
<td>0.002</td>
</tr>
<tr>
<td>2</td>
<td>0.001</td>
<td>29.94</td>
<td>1.00</td>
<td>0.004</td>
</tr>
<tr>
<td>3</td>
<td>0.010</td>
<td>60.25</td>
<td>1.29</td>
<td>0.170</td>
</tr>
<tr>
<td>4</td>
<td>0.096</td>
<td>78.77</td>
<td>2.07</td>
<td>0.819</td>
</tr>
<tr>
<td>5</td>
<td>0.886</td>
<td>85.75</td>
<td>9.94</td>
<td>0.005</td>
</tr>
<tr>
<td>6</td>
<td>0.004</td>
<td>93.80</td>
<td>1.02</td>
<td>0.000</td>
</tr>
<tr>
<td>7</td>
<td>0.001</td>
<td>101.8</td>
<td>1.01</td>
<td>0.000</td>
</tr>
<tr>
<td>8</td>
<td>0.000</td>
<td>109.7</td>
<td>1.00</td>
<td>0.000</td>
</tr>
<tr>
<td>9</td>
<td>0.000</td>
<td>117.8</td>
<td>1.00</td>
<td>0.000</td>
</tr>
<tr>
<td>10</td>
<td>0.000</td>
<td>125.8</td>
<td>1.01</td>
<td>0.000</td>
</tr>
</tbody>
</table>

3.2. Example 2

In this example we simulated univariate normal samples for high ($n = 1000, N(85, 100)$) and low ($n = 500, N(65, 49)$) diagnostic groups to assess the possibility of overfitting by the MDP approach. As in Example 3.1, we specified the upper bound for the number of mixtures in each group as 10 and used the same priors for $\alpha$ and component parameters, even though we knew otherwise. Table II shows the posterior means of the mixing weights and parameters $\mu_i$’s and $\sigma_i$’s in each group. In the high group, the majority (88.6 per cent) of the posterior simulations were drawn approximately from a $N(85.75, 98.80)$ and the rest were simulated from distributions associated with very small mixing weights. In the low group, one dominating component (81.9 per cent) and one low frequency component (17 per cent) were identified with the rest being negligible. Figure 4 presents the posterior predictive densities obtained using three different approaches all of which reproduced the true distributions very closely. In Figure 5, the estimates of TPR, FPR, and ROC curves are also similar for all three approaches. The estimated AUC is 0.9453 with a 95 per cent CI of (0.9347, 0.9555). This example shows that overfitting is not a major issue for the MDP model.

3.3. Real data: Great Smoky Mountains Study

The Great Smoky Mountains Study (GSMS) [1] is a population-based study of youth from western North Carolina counties that entered the study at ages 9, 11, and 13. Using data on students attending schools in 11 counties, an initial sample of 4500 youth was selected using a household equal probability design. A short screen based on the Child Behavior Check List (CBCL) [29] was administered to all 4500 children. Twenty-five percent of the sample ($n = 1125$) screened high and 75 per cent ($n = 3375$) screened low. Due to budget constraints, the sample was reduced to $n = 1462$ by retaining all subjects who screened high ($n = 1125$) and randomly selecting 10 per cent of the subjects who screened low ($n = 337$). Subjects subsequently completed the Child and Adolescent Psychiatric Assessment (CAPA), an interview-based diagnostic tool for arriving at diagnoses based on DSM-IV criteria [30, 31]. There are, therefore, screening scores for 4500 youth and diagnoses for 1462 youth, with diagnoses for the remaining $n = 3038$ subjects being missing at random. The CAPA provides information
Figure 4. Application to univariate normal diagnostic populations: MDP, Normal models, and KDE.

Figure 5. Posterior estimates of TPR, FPR and ROC based on the MCMC simulations for univariate Normal diagnostic populations: MDP, Normal models, and kernel density estimator (KDE).

about multiple diagnoses, but in this example the diagnosis of interest is depression. For the purpose of this example, we chose a random subset of the subjects in the high-screen group with the same selection probability of 10 per cent as in the low screen group, so that the resulting sample of \( n = 337 + 112 = 449 \) is a random sample from the entire population.

In our previous examples, we implicitly assumed the existence of a gold standard diagnostic test by which the population of interest was correctly classified into mutually exclusive, diseased and non-diseased groups. In this example, there is no gold standard. Diagnoses derived by using the diagnostic interview used in GSMS are based on diagnostic criteria set out in the DSM-IV and are the diagnostic standard in mental health, but not a perfect gold standard. Our purpose here is to compare semi-parametric ROC analyses obtained under the assumptions of the existence and absence of a gold standard for diagnosing depression, using the CBCL scores as a screening tool.

### 3.3.1. Modelling without a gold standard.

The use of latent class models has been a major trend in adjusting for imperfect standards. See for example Reference [32] for a random-effects latent-class model, and Yang and Becker [33], for a marginal latent-class model. Following the current trend, we introduce a binary variable \( V_i \) to represent the true but latent diagnostic status of depression for the \( i \)th subject. To relate the imperfect standard \( D_i \) to \( V_i \) we consider two measurement error structures: a classical error model and a Berkson error model [34]. In the classical error model, \( D_i \) is conditioned on \( V_i \) whereas \( V_i \) is conditioned on \( D_i \) in the Berkson error model.

#### Classical error model:

\[
D_i | V_i \sim \text{Bernoulli}(p_i) \\
\text{Logit}(p_i) = \beta_0 + \beta_1 V_i \\
V_i \sim \text{Bernoulli}(q_i) \\
q_i \sim \text{Beta}(a,b)
\]

#### Berkson error model:

\[
V_i | D_i \sim \text{Bernoulli}(p_i) \\
\text{Logit}(p_i) = \beta_0 + \beta_1 D_i
\]

The intercept and slope parameters \( \beta_0 \) and \( \beta_1 \) can be fixed or assigned prior distributions to reflect uncertainty about the association between \( D_i \) and latent \( V_i \). The hierarchical models for \( X \) (GSMS screen scores based on the CBCL) are assumed to have the same structure as in equation (A1) in Appendix A, after conditioning on the true depression diagnosis \( V_i \) and imperfect depression diagnosis \( D_i \):

\[
X_i | V_i, D_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_{K_i|V_i,D_i}, \tau_{K_i|V_i,D_i}), \quad i = 1, \ldots, n
\]

For a randomly selected triplet \([X = x, D, V]\) (in this example they are, respectively, a future CBCL screen score, a future imperfect depression diagnosis, and an unknown but true...
depression diagnosis), the posterior predictive density \( f_\nu(x \mid X_D) \) is obtained as analogous to (A7)

\[
f_\nu(x \mid X_D) = E \left[ \sum_{k=0}^{C} w_{k|V,D} f(x \mid \theta_{k|V,D}^{(i)}) \right] \approx \frac{1}{T} \sum_{i=1}^{T} \left[ \sum_{k=0}^{C} w_{k|V,D}^{(i)} f(x \mid \theta_{k|V,D}^{(i)}) \right]
\]

where \( f(x \mid \theta_{k|V,D}^{(i)}) \) is the normal density for \( x \mid \theta_{k|V,D}^{(i)} \sim N(\mu_{k|V,D}, \tau_{k|V,D}) \) evaluated at the draw \( \theta_{k|V,D}^{(i)} \). The predictive distribution functions are obtained by integrating the mixture density above for each value of the cut-off \( c \) as in (A10) and (A11). For the purpose of illustration, the prior distribution of \( \{ \mu_{k|V,D} \} \) is assumed to follow the ordered-normal priors described by independently and normally distributed increments with mean 0 and precision 0.10 (i.e. variance of 10), and \( \tau_{k|V,D} \) is assumed to follow a Gamma(0.01, 0.1) prior, for each level of \( [V,D] \). Note that in the above formulation \( D_i \) is fixed, but \( V_i \) is an unknown latent variable which can be simulated from its conditional posterior as an additional step to the Gibbs sampler described in Appendix B.

**Classical error model:**

\[
V_i \mid D_i, X_0, X_i, \{ \mu_{k|V,D}, \tau_{k|V,D} \}, K_i, \beta_0, \beta_1 \sim Bernoulli(q_i^*)
\]

**Berkson error model:**

\[
V_i \mid D_i, X_0, X_i, \{ \mu_{k|V,D}, \tau_{k|V,D} \}, K_i, \beta_0, \beta_1 \sim Bernoulli(p_i^*)
\]

where \( q_i^* \) is defined as, up to a constant of proportionality, for \( i = 1, \ldots, n \)

\[
q_i^* \propto [X_i \mid V_i, D_i, K_i, \{ \mu_{k|V,D}, \tau_{k|V,D} \}][K_i \mid D_i][D_i \mid V_i][V_i]
\]

and \( p_i^* \) is defined as, up to a constant of proportionality

\[
p_i^* \propto [X_i \mid V_i, D_i, K_i, \{ \mu_{k|V,D}, \tau_{k|V,D} \}][K_i \mid D_i][V_i \mid D_i]
\]

both being weighted averages of normal densities

\[
\phi(x_i; \mu_{k|V} = 0, D_i, \tau_{k|V} = 0, D_i) \quad \text{and} \quad \phi(x_i; \mu_{k|V} = 1, D_i, \tau_{k|V} = 1, D_i)
\]

### 3.3.2. Application to GSMS data

We first treated the current depression diagnosis as a gold standard and fit the MDP model as before to obtain the posterior predictive estimate of the ROC curve. We then applied the extended MDP model (extended by \( \{ V_i \} \)) described above under the assumption of an imperfect gold standard and considered both the classical error model in (3) and the Berkson error model in (4). We specified an upper bound for the number of mixtures as \( C = 5 \) for each pair of groups \( \{ V = 1, D = 1 \}, \{ V = 1, D = 0 \}, \{ V = 0, D = 1 \} \) and \( \{ V = 0, D = 0 \} \). The parameters, \( \beta_0 \) and \( \beta_1 \) in (3) and (4) each were assumed to be independently distributed with a proper but non-informative normal prior \( N(0, 10) \). In the classical error model setup in (3), the latent true status \( V_i \) was assumed to follow Bernoulli(q_i), where \( q_i \sim \text{Beta}(1, 1) \).

The posterior predictive densities obtained under both assumptions, with and without a gold standard, are presented in Figure 6. We note that both predictive estimates based on the MDP (with and without the gold standard) fit the data quite well, and are very similar to
Figure 6. Posterior predictive densities of $[X \mid D]$ (first two) and $[X \mid V, D]$ (bottom four) under MDP comparing models with and without gold standard (Berkson type model, non-Berkson type model) (overlayed on the histogram of actual data).

each other. As for the models without the gold standard, the Berkson error model seemed to show a marginally better fit to the data than the classical error model. We also illustrate the conditional predictive densities $[X \mid V, D]$ in Figure 6 (bottom four figures) under both classical and Berkson error assumptions. Notice that the predictive distributions $[X \mid D = 0, V]$ and $[X \mid D = 1, V]$ are different for $D = 0$ and $D = 1$ for all $V$, showing evidence against non-differential misclassification, i.e. we have $[X \mid D, V] \neq [X \mid V]$ for this data set.

Figure 7 presents the predictive estimates of TPR, FPR, and the ROC curves obtained under both assumptions. Figure 7 only presents the ROC obtained from the Berkson error model because the ROC obtained from classical error model was almost identical. The posterior distributions of the intercept and slope parameters described in equations (3) and (4) are shown in Figure 8. The posterior means (variances) were $-3.46$ (1.34) for $\beta_0$, and $2.84$ (1.74) for $\beta_1$, for the classical error model, and $-0.11$ (1.89), $3.15$ (6.18) for the Berkson error model, respectively, under the non-informative but proper priors. The posterior distributions of $\beta_0$ and $\beta_1$ are informative even though relatively weak priors were used in the examples. That is, there is learning about $\beta_0$ and $\beta_1$ regardless of the measurement error model entertained, and they are identifiable.
Figure 7. MDP estimates of TPR, FPR and ROC curves for comparing models with and without gold standard (Berkson model).

Figure 8. Prior and posterior probabilities of $\beta_0$, $\beta_1$ comparing Berkson and classical error models.
The ROC curves are very close, although substantial uncertainty remains to be explained for the ROC curve obtained under the imperfect gold standard assumption. This is largely due to the uncertainty in the unknown true state of depression status, and can be overcome using a strong (highly informative) discrete prior for $V_i$. As evidenced in Figure 7, the semi-parametric model (A1) performs quite robustly even if the true depression diagnosis is not a gold standard, and a latent class extension of the MDP model seems unnecessary in practice.

4. CONCLUSIONS

In this paper, we developed an efficient semi-parametric approach for estimating ROC curves and AUC. Using a block Gibbs sampling approach, posterior distributions of the MDP model parameters can be efficiently computed, thus providing direct estimates of posterior uncertainty in the estimation of the ROC curves, sensitivity (TPR) and specificity (1-FPR) of a diagnostic test. The inferences on ROC curves based on both the MDP and KDE approaches were shown to be very similar. As illustrated in the GSMS example, the MDP provides robust estimation of ROC curves even under the assumption of imperfect diagnosis. We are currently developing an ROC approach based on mixtures of multivariate normal distributions to model multiple (correlated) diagnostic and screen tests.

APPENDIX A

A.1. Mixtures of Dirichlet process priors

Let $\mathbf{X}$ denote the (continuous) outcome of a test device designed to detect the true signal $D=1$ and the noise signal $D=0$. For simplicity we suppress the dependency on $D$. Consider the following hierarchical model [12, 24, 35]:

$$
\mathbf{X}_i|\theta, \mathbf{K} \sim^i N(\mu_{K_i}, \tau_{K_i}), \quad i = 1, \ldots, n
$$

$$
K_i|\mathbf{w} \sim \sum_{k=1}^{C} w_k \delta_{\phi_k}(\cdot) = \text{Multinomial}(1, \mathbf{w}_C)
$$

and

$$
\theta_k = (\mu_k, \tau_k) \sim G_0, \quad k = 1, \ldots, C
$$

(A1)

where $\tau_k = 1/\sigma_k^2$, $\delta_{\phi}(b)$ is the unit mass at $a = b$, $\theta = (\theta_1, \ldots, \theta_n)$, $\mathbf{K} = (K_1, \ldots, K_n) \in \{1, \ldots, C\}^n$ and the weights $\mathbf{w} = (w_1, \ldots, w_C)$ are determined by the stick-breaking algorithm

$$
w_1 = R_1
$$

and

$$
w_k = (1 - R_1)(1 - R_2) \ldots (1 - R_{k-1}) R_k, \quad k = 2, \ldots, C - 1
$$

(A1*)
with \( \{w_k\} \) being an iid sequence of Beta(1,\( \alpha \))'s, \( R_C = 1 \) to make sure that \( w_k \)'s add up to unity. Ishwaran and James [24] show that the hierarchical model described in (A1) arise from an approximation

\[
G_C(\cdot) = \sum_{k=1}^{C} w_k \delta_{\theta_k}(\cdot)
\]

to a Dirichlet Process (DP)

\[
G(\cdot) = \sum_{k=1}^{\infty} w_k \delta_{\theta_k}(\cdot), \quad \theta_k \sim G_0 \quad \text{for } k \geq 1
\]

with base measure \( \propto G_0 \) being independent of the weights \( w = (w_1, \ldots, w_\infty) \). Moreover, for sufficiently large \( C \), i.e. \( C \to \infty \) the \( G_C \) converges to \( G \) almost surely [23, 36]. The discrete random variables in \( \mathbf{K} \) are classification variables, each describing common elements in the parameter vector \( \theta \), which occurs with non-zero probability due to the discreteness of \( G_C \) (and \( G \) in the limit); \( k_i = k_j = j \) if the observations \( x_i \) and \( x_j \) share the same parameter \( \theta_i = \theta_j = \theta_j \). For the purposes of this paper, we choose a conjugate normal–gamma baseline prior \( G_0 \) for the components of \( \theta \)'s independently

\[
\mu_i \overset{\text{ind}}{\sim} \mathcal{N}(\xi, \varphi)
\]
\[
\tau_i \overset{\text{ind}}{\sim} \text{Gamma}(u, v)
\]

so that \( dG_0(\theta_i) = \phi(\mu_i | \xi, \varphi) \Gamma(\tau_i | u, v) d\mu_i d\tau_i \), where \( \phi \) and \( \Gamma \) denote the densities of a normal and a gamma variate, and the mean \( \xi \), precision \( \varphi, u, \) and \( v \) are assumed known. If the concentration parameter is also unknown, a Gamma prior is a common choice. It is also possible to consider non-conjugate baseline prior distribution \( G_0 \), although at the expense of additional computational complexity.

A.2. ROC analysis using mixtures of DPP

The key element in our MDP specification for ROC analyses is the posterior predictive density of a future score \( X_{n+1} = x \) given the data \( \mathbf{X} \)

\[
f(x | \mathbf{X}) = \int \cdots \int f(x | \theta_1, \ldots, \theta_n) g(\theta_1, \ldots, \theta_n | x_1, \ldots, x_n) d\theta_1 \ldots d\theta_n
\]

\[= E\{ f(x | \theta_1, \ldots, \theta_n) \} \quad \text{(A2)}
\]

where the expectation is taken with respect to the posterior distribution \( [\theta, \mathbf{K} | \mathbf{X}] \) under finite-mixture MDP model \( G_C \). The first integrand in (A2) is the finite mixture [20, 24]

\[
f(x | \theta_1, \ldots, \theta_n) = \sum_{k=1}^{C} \tilde{w}_k f(x | \tilde{\theta}_k) \quad \text{(A3)}
\]

where the parameters \( (\tilde{w}_k, \tilde{\theta}_k = (\tilde{\mu}_k, \tilde{\tau}_k) \) are random draws taken from the probability measure \( G_C \), and \( f(x | \tilde{\theta}_k) \) is the corresponding normal density. Notice that (A3) has the form of
a kernel density estimate with normal kernels \( f(x \mid \hat{\theta}_k) \) as in the KDE approach. The second integrand in (A2) is the posterior distribution of the vector \( \theta \) under \( G_C \), conditional on the data \( X = \{x_1, \ldots, x_n\} \). Exact computation of (A2) is difficult even for small sample sizes. However, a Gibbs sampler can be developed for generating samples from the posterior under \( G_C \). Consequently, the predictive density in (A2) can be approximated by a Monte Carlo average

\[
\hat{f}(x \mid X) = \frac{1}{T} \sum_{t=1}^{T} \left[ \sum_{k=1}^{C} w^{(t)}_k f(x \mid \theta^{(t)}_k) \right]
\]

(A4)

where \( \{(w^{(t)}_k, \theta^{(t)}_k)\}_{t=1}^{T} \) is a Monte Carlo sample of size \( T \) from the posterior \( [\theta, w, K \mid X] \) under \( G_C \). The Monte Carlo approximation in (A4) to posterior predictive density in (A2) constitutes the basis of our approach which is used to estimate the predictive densities under the non-diseased \( (D = 0) \) and diseased \( (D = 1) \) populations, and sensitivities and specificities, and ROC curve, respectively.

For the non-diseased and diseased populations, \( D = 0 \) and \( D = 1 \), denote the posterior predictive distribution functions obtained using the mixtures of MDP approach described above as \( F_D(x \mid X_D) \), and the posterior predictive densities as \( f_D(x \mid X_D) \) where \( X_D \) is the vector of the independent observations (screen scores), each following a hierarchical mixture model as described in equation (A1). Let \( Y \) be a decision variable taking a value of 1 if a randomly selected subject’s \( X \) exceeds a cut-off \( c \) and a value of 0, otherwise. Then, the predictive TPR and FPR are defined as

\[
\lambda(c \mid X_1) = \Pr(Y = 1 \mid X_1) = \Pr(X \geq c \mid X_1) = 1 - F_1(c \mid X_1)
\]

(A5)

and

\[
\gamma(c \mid X_0) = \Pr(Y = 1 \mid X_0) = \Pr(X \geq c \mid X_0) = 1 - F_0(c \mid X_0)
\]

(A6)

Since

\[
F_D(c \mid X_D) = \int_{-\infty}^{c} f_D(x \mid X_D) \, dx
\]

\[
= \int_{-\infty}^{c} \left\{ \int f_D(x \mid \theta_D)g(\theta_D \mid X_D) \, d\theta_D \right\} \, dx
\]

(A7)

where \( \theta_D \) denotes the parameter vector size for group \( D \), interchanging the order of integration gives

\[
F_D(c \mid X_D) = \int F_D(c \mid \theta_D)g(\theta_D \mid X_D) \, d\theta_D = E[F_D(c \mid \theta_D)]
\]

(A8)

where the expectation is taken with respect to the posterior distribution of \( [\theta_D, w, K_D \mid X_D] \) under \( G_{D_k} \). By equation (A3)

\[
F_D(c \mid \theta_D) = \sum_{k=1}^{C_D} w_{Dk} F_D(c \mid \theta_{Dk})
\]

(A9)
so that
\[
\lambda(c \mid X_1) = E \left( \frac{C_j}{\sum_k w_{1k}[1 - F_1(c \mid \theta_{1k})]} \right) = E \left[ \sum_{k=1}^{C_j} w_{1k} \lambda_k(c \mid \theta_{1k}) \right]
\] (A10)
and
\[
\gamma(c \mid X_0) = E \left( \frac{C_k}{\sum_k w_{0k}[1 - F_0(c \mid \theta_{0k})]} \right) = E \left[ \sum_{k=1}^{C_k} w_{0k} \gamma_k(c \mid \theta_{0k}) \right]
\] (A11)
which together describe posterior predictive ROC curve by varying the cut-off \(c\). Again, expectations \(E\) are with respect to posterior distributions induced by \(G_0\) for \(D=0\) and \(D=1\). Equations (A10) and (A11) are the predictive TPR and FPR described as finite (normal) mixtures of component TPRs and FPRs averaged over the posterior distributions of the unknown parameters in the models for \(D=0\) and \(D=1\) populations. Solving (A10) and (A11) numerically for \(c\), an equivalent representation of ROC curve can be obtained in terms quantiles
\[F_1^{-1}(q_1 \mid X_1) = F_0^{-1}(q_0 \mid X_0)\]
where \(F_1(\cdot \mid X_1)\) and \(F_0(\cdot \mid X_0)\) are as defined in (A8) and (A9). We can also compute the area under ROC curve, AUC, as
\[
AUC = \int_{-\infty}^{\infty} \left[ \int_{-\infty}^{c} dF_0(x \mid X_0) \right] dF_1(x \mid X_1) = E \left[ \sum_{k=1}^{C_k} \sum_{l=1}^{C_l} w_{0k} w_{1l} \left\{ \int_{-\infty}^{\infty} F_0(c \mid \theta_{0k}) F_1(c \mid \theta_{1l}) dc \right\} \right]
\]
where the expectation is with respect to the joint posterior \([\theta_0, w_0, K_0 \mid X_0] \times [\theta_1, w_1, K_1 \mid X_1]\). Under the normal mixture model (A1) for \(D=0\) and \(D=1\) populations, where \(\theta_{0k} = (\mu_{0k}, \tau_{0k})\) and \(\theta_{1l} = (\mu_{1l}, \tau_{1l})\), the integral on the right-hand side has the closed analytical form, as in the binormal model
\[
\int_{-\infty}^{\infty} F_0(c \mid \theta_{0k}) F_1(c \mid \theta_{1l}) dc = \Phi(b_{kl} / \sqrt{1 + a_{kl}^2}) \overset{\text{def}}{=} AUC(k,l)
\]
where \(\Phi\) denotes the standard normal distribution function, \(b_{kl} = \tau_{1l}^{1/2} (\mu_{1l} - \mu_{0k})\), and \(a_{kl} = [\tau_{1l}/\tau_{0k}]^{1/2}\), so that
\[
AUC = E \left[ \sum_{k=1}^{C_k} \sum_{l=1}^{C_l} w_{0k} w_{1l} AUC(k,l) \right]
\] (A12)
a finite mixture of AUCs with doubly mixing weights \([w_{0k} w_{1l}]\) summing to unity over indices \(k\) and \(l\). Since the posterior distributions \([\theta_0, w_0, K_0 \mid X_0]\) are not available in closed form, the MCMC (e.g., Gibbs sampler) sampled values of the parameters \([b_{kl}, a_{kl}]\) and \([w_{0k}, w_{1l}]\) obtained from \([G_0, G_1]\) can be used to make statistical inferences about the ROC curves described by (A10) and (A11), and AUC in (A12), by replacing the high-dimensional posterior expectations with the Monte Carlo averages as in (A4) above. See Appendix B for a Gibbs sampling approach for posterior computations used in approximating (A10), (A11),
and (A12), and Appendix C for the WinBUGS codes. Posterior simulations of TPR, FPR, ROC curve and AUC can be done directly within WinBUGS, or outside of WinBUGS using spreadsheets or other convenient programs by storing a few thousand Gibbs realizations of the parameters after convergence has been established.

APPENDIX B: GIBBS SAMPLER FOR POSTERIOR COMPUTATIONS

The early implementations of MDP models relied on Polya urn Gibbs sampling algorithms developed by Escobar [37] and later followed by various modifications to speed up the computations, see References [14, 21, 38]. For the purpose of this paper, we adopt a blocked Gibbs sampling approach described in Ishwaran and James [24]. This method relies on the hierarchical model and finite approximation in equation (A1) and has been shown to have good mixing properties [39]. For clarity we drop the population index $D$ below, with the understanding that the Gibbs sampling algorithm is being applied separately to the $D=0$ and $D=1$ groups. From equation (A1), and using the priors specified therein, the joint posterior distribution $[\theta, w_c, K | X]$ can be obtained using the Bayes theorem as

$$
[\theta, w_c, K | X] \propto \prod_{i=1}^{n} \{ N(x_i|\mu_{K_i}, \tau_{K_i}) \times N(\mu_{K_i}|\xi, \varphi) \times \text{Gamma}(\tau_{K_i}|\alpha, \beta) \}
\times \prod_{i=1}^{n} \{ \text{Multinomial}(K_i|1,w_C) \} \times \pi(w_c = w_c(R_{C-1}))
$$

(B1)

where $\pi$ is the prior distribution of $w_C$ induced by the stick-breaking algorithm described in (A1*) with independent beta priors $\{ \text{Beta}(R_k|1,\alpha) \}_{k=1}^{C-1}$. For illustration purposes we assumed below that the concentration parameter $\alpha$ is assumed fixed in advance. We discussed the choice of priors for, and posterior simulation of, $\alpha$ below. The joint posterior distribution in (B1) is not available in closed analytical form, however, it leads to the following full conditional distributions that are available in closed form:

Con1: $[\theta | w_c, K, X]$

$$
= [\theta | K, X] \propto \prod_{i=1}^{n} \{ N(x_i|\mu_{K_i}, \tau_{K_i}) \times N(\mu_{K_i}|\xi, \varphi) \times \text{Gamma}(\tau_{K_i}|\alpha, \beta) \}
$$

Con2: $[K | w_c, \theta, X]$

$$
= [K | w_c, \theta, X] \propto \prod_{i=1}^{n} \{ N(x_i|\mu_{K_i}, \tau_{K_i}) \times \text{Multinomial}(K_i|1,w_C) \}
$$

Con3: $[w_c | \theta, K, X]$

$$
= [w_c | K] \propto \prod_{i=1}^{n} \{ \text{Multinomial}(K_i|1,w_C) \} \times \pi(w_c = w_c(R_{C-1}))
$$
Let \( \{K^*_1, \ldots, K^*_n\} \) denote the current \( C^* = m \leq n \) distinct values of \( K \), and define the number of distinct elements in \( K \) as \( n_j = \#\{i: K_i = j\} \) for any \( j \in \{K^*_1, \ldots, K^*_n\} \). Then, the Con1 can further be decomposed as, with \( \tau = (\tau_1, \ldots, \tau_n) \), and \( \mu = (\mu_1, \ldots, \mu_n) \)

\[
\mu_j | \tau, K, \mathbf{X} \sim \begin{cases} 
N(\mu_j | \mu^*_j, \tau^*_j) & \text{for } j \in \{K^*_1, \ldots, K^*_n\} \\
N(\mu_j | \mu^*_j, \varphi) & \text{for } j \in K \setminus \{K^*_1, \ldots, K^*_n\}
\end{cases}
\]

and

\[
\tau_j | \mu, K, \mathbf{X} \sim \begin{cases} 
\text{Gamma}(\tau_j | u^*_j, v^*_j) & \text{for } j \in \{K^*_1, \ldots, K^*_n\} \\
\text{Gamma}(\mu_j | u, v) & \text{for } j \in K \setminus \{K^*_1, \ldots, K^*_n\}
\end{cases}
\]

where \( \bar{x}_j = n_j^{-1} \left( \sum_{i:K_i = j} x_i \right) \), \( \mu^*_j = \tau_j^{-1} (\tau_j^{-1} n_j \bar{x}_j + \varphi^{-1} \bar{x}) \), \( \tau^*_j = (n_j \tau_j^{-1} + \varphi^{-1}) \), \( u^*_j = u + n_j/2 \), and \( v^*_j = v + \sum_{i:K_i = j} (x_i - \mu_j)^2/2 \). Similarly, Con2 can be written as

\[
K_i | w_{C,i}, 0, \mathbf{X} \sim \text{Multinomial}(1, \mathbf{w}_{C,i}), \quad i = 1, \ldots, n
\]

where \( w_{C,i} = (w_{1,i}, \ldots, w_{C,i}) \), \( w_{C,i} \propto \exp(- (x_i - \mu_i)^2/2) \), \( \sum_{i=1}^n w_{C,i} = 1 \).

In Con3, the conditional posterior for \( C \) is obtained from the conjugate update for the stick-breaking algorithm

\[
w_1 = R^*_1
\]

and

\[
w_k = (1 - R^*_1)(1 - R^*_2) \cdots (1 - R^*_{k-1})R^*_k, \quad k = 2, \ldots, C - 1
\]

where \( R^*_k \sim \text{Beta}(1 + n_k, \alpha + \sum_{l=k+1}^C n_l) \), \( n_l = \#\{i: K_i = l\} \), \( k = 1, \ldots, C - 1 \). Thus, starting from an initial state \((\theta^{(0)}, w^{(0)}_{C}, K^{(0)})\) we can sequentially simulate \((\theta^{(l)}, w^{(l)}_{C}, K^{(l)})\) from the conditionals Con1, Con2 and Con3. The approximate draws of \((\theta^{(l)}, w^{(l)}_{C}, K^{(l)})\) from the posterior distribution \( [\theta, w, K | \mathbf{X}] \) are then used to compute the posterior predictive density described in (A4), and posterior predictive distributions of ROC and AUC. The sampled ROC curves provide information relevant to assessing posterior uncertainty about the ROC estimate, i.e. the posterior mean of the sampled ROC curves, conditional on the current data available from \( D = 0 \) and \( D = 1 \) groups, respectively.

B.1. Prior elicitation and posterior simulation of \( \alpha \)

There are several ways of choosing a suitable prior for \( \alpha \). Ishwaran et al. [39] and Ishwaran and James [24] show that, in the sup metric, the difference between the marginal density of \( \mathbf{X} = \{x_1, \ldots, x_n\} \) under \( G \) and the marginal density of \( \mathbf{X} \) under finite approximation \( G_C \) has an error bound \( B = 4n \exp(-(C - 1)/\alpha) \), which may be used to choose the truncation limit \( C \) (upper bound for the number of mixtures), and concentration parameter \( \alpha \). Alternatively, one can use the approximations suggested by Liu [25], who show that the conditional prior mean and variance of the number of distinct \( C^* \) components are

\[
E(C^* | \alpha, n) \approx \alpha \log(1 + n/\alpha)
\]

\[\text{(B2)}\]
and

$$\text{Var}(C^*|x, n) \approx \alpha[\log(1 + n/\alpha) - 1]$$

(B3)

For fixed values of $\alpha$, we can use equations (B2) and (B3) to assess prior uncertainties about the number of distinct clusters. From (B3), we must have $\log(1 + n/\alpha) > 1$ so that $\alpha < n/(e - 1)$, where $e = 2.7182$ so that $E(C^*|x, n) > \alpha$ for all $\alpha > 0$. For example, for $n = 100$ and $c = 1$; $E(C^*|x, n) \approx 5$, $\text{Var}(C^*|x, n) \approx 4$ and for $\alpha = 58$; $E(C^*|x, n) \approx 58.12$, $\text{Var}(C^*|x, n) \approx 0.12$. Thus, one can choose $\alpha$ to match the prior moments in (B2) and (B3) to pre-specified values reflecting prior beliefs/information about the number of distinct clusters.

If $\alpha$ is treated as a random variable distributed according to a Gamma($a_0, b_0$) prior, then Taylor series expansions of (B2) and (B3) around the prior mean $E(\alpha) = a_0/b_0$ yield

$$E(C^*) \approx (a_0/b_0) \log(1 + b_0n/a_0)$$

(B4)

and

$$\text{Var}(C^*) \approx E(C^*) - (a_0/b_0) + (a_0/b_0^2)[\log(1 + b_0n/a_0) - nb_0/(a_0 + nb_0)]^2$$

(B5)

two (non-linear) equations in two unknowns $a_0$ and $b_0$, which can be solved numerically to elicit prior information about $\alpha$. The posterior distribution of $\alpha$ can be simulated using the conditional posterior in obtained from (B1) and Con3 above, as

$$\text{Con4: } [\alpha|w, 0, K, X] = [\alpha|w_C] \propto \pi(w_C = w_C(R^*_C)) \text{Gamma}(a_0, b_0)$$

where $\pi(w_C = w_C(R^*_C))$ is the conditional density of the weights induced by $R^*_1, \ldots, R^*_C$ as described in Con3 above. It can be shown that (see Reference [40]) Con4 leads to the conditional posterior

$$\alpha|w_C \sim \text{Gamma}(C + a_0 - 1, b_0 - \sum_{c=1}^C \log(1 - R^*_c))$$

which can be sampled easily.

APPENDIX C: WINBUGS CODE

C.1. WinBUGS code for Section 3.1: Example with simulated data (Gold standard test)

Note: To be fitted separately to $D = 0$ and $D = 1$ groups.

```r
model
{
  for (i in 1:n)
  X[i] ~ dnorm(mu[i], tau[i]) #screen score for subject 'i'

  for (i in 1:n)
  K[i] ~ dcat(p[1:C]) #represents membership of 'C' mixtures of normals

  mu[i] <- theta[K[i]] #parameters corresponding to each of mixtures of normals
  tau[i] <- delta[K[i]]
}
```

Copyright © 2006 John Wiley & Sons, Ltd.

for (j in 1:C) {
    SC[i,j] <- equals(j,K[i])  } #either one or zero representing membership of each
group, to estimate the number of mixtures

#Precision Parameter
alpha ~ dgamma(3.37, 5.28) #precision of Dirichlet, optimal shape and scale parameters
determined by the mean and the variance of clusters

#Specified hyper parameters
mu.theta <- 0
tau.theta <- 1
u.tau <- 0.01
v.tau <- 0.1

#Constructive MDP using Stick Breaking Algorithm (See Reference [41, Example 6.27])

#Ordering Theta
theta[i,1] ~ dnorm(mu.theta, tau.theta)
for (i in 2:C) {theta[i] ~ dnorm(-theta[i-1]+inc[i-1])}
for (i in 1:C-1) {inc[i] ~ dnorm(0, 0.01) if (i)}

# Estimate the total number of clusters
for (j in 1:C) {S,SC[i] <- step(sum(SC[,j])-1)}
Cluster <- sum(S,SC[])
}

C.2. WinBUGS code for Section 3.2: Great Smoky Mountains Study (imperfect gold
standard test for classical error model)

model
{
for (i in 1:n) {
  X[i] ~ dnorm(mu[i], tau)  #imperfect screen score for subject i
  D[i] ~ dbern(p[i])  # a binary latent variable representing the true and unknown
diagnostic status for subject i, where 0 = non-diseased, 1 = diseased
  logit(p[i]) <- b1 + b2*V[i]  #logit link between true diagnosis V and imperfect
diagnostic test D
  v[i] ~ dbeta(1,1)
  q[i] ~ dgamma(0.01, 0.1)  #prior for precision of screen score
  b1 ~ dnorm(0, 0.1)  #specified prior for intercept of logit link
  b2 ~ dnorm(0, 0.1)  #specified prior for slope of logit link
  #indicator for non-diseased and diseased
  for (i in 1:N) {
    I00[i] <- equals(0, V[i])*equals(0, D[i])  #indicator for those with v=0, D=0
    I01[i] <- equals(0, V[i])*equals(1, D[i])  #indicator for those with v=0, D=1
    II0[i] <- equals(1, V[i])*equals(0, D[i])  #indicator for those with v=1, D=0
    II1[i] <- equals(1, V[i])*equals(1, D[i])  #indicator for those with v=1, D=1
  }
  #mixture grouping variable
  for (i in 1:N) {
    K00[i] ~ deat(p00[1:C])
    K01[i] ~ deat(p01[1:C])
    K10[i] ~ deat(p10[1:C])
  }
K11[i] ~ dcat(p11[i])

#mixtures on the mean of screen score separately for (v=0,D=1), (v=0,D=1), (v=1,D=0),
(v=1,D=1) for (i in 1:N)
mu[i] <- theta00[K00[i]]*100[i] + theta01[K01[i]]*101[i] + theta10[K10[i]]*110[i]
+ theta11[K11[i]]*111[i]

#Precision Parameter
alpha ~ dgamma(3.6,3.1) #precision of Dirichlet, specified values found by optimization for the
mean and the variation of the number of cluster

#specified hyper parameters
mu,theta <- 0
tau,theta <- 0.1

#Constructive MDP using stick breaking algorithm
p00[1] <- r00[1]
for (j in 2:C) { p00[j] <- r00[j] *(1-r00[j-1])*(p00[j-1]/r00[j-1])
for (k in 1:C-1) { r00[k] ~ dbeta(1, alpha) }
r00[C] <- 1 #ensure sum to 1
}

p01[1] <- r01[1]
for (j in 2:C) { p01[j] <- r01[j] *(1-r01[j-1])*(p01[j-1]/r01[j-1])
for (k in 1:C-1) { r01[k] ~ dbeta(1, alpha) }
r01[C] <- 1 #ensure sum to 1
}

p10[1] <- r10[1]
for (j in 2:C) { p10[j] <- r10[j] *(1-r10[j-1])*(p10[j-1]/r10[j-1])
for (k in 1:C-1) { r10[k] ~ dbeta(1, alpha) }
r10[C] <- 1 #ensure sum to 1
}

for (j in 2:C) { p11[j] <- r11[j] *(1-r11[j-1])*(p11[j-1]/r11[j-1])
for (k in 1:C-1) { r11[k] ~ dbeta(1, alpha) }
r11[C] <- 1 #ensure sum to 1
}

# Ordering Theta
theta00[1] ~ dnorm(mu,theta,tau,theta)
theta01[1] ~ dnorm(mu,theta,tau,theta)
theta10[1] ~ dnorm(mu,theta,tau,theta)
theta11[1] ~ dnorm(mu,theta,tau,theta)
for (i in 2:C){
theta00[i] <- theta00[i-1] + inc00[i-1]
theta01[i] <- theta01[i-1] + inc01[i-1]
theta10[i] <- theta10[i-1] + inc10[i-1]
theta11[i] <- theta11[i-1] + inc11[i-1]
for (i in 1:C-1) {
inc00[i] ~ dnorm(0,0.1)[0,0)
inc01[i] ~ dnorm(0,0.1)[0,0)
inc10[i] ~ dnorm(0,0.1)[0,0)
inc11[i] ~ dnorm(0,0.1)[0,0)
}

ACKNOWLEDGEMENTS

This research was supported by Grants from the National Institute of Health (MH-66497).

REFERENCES