# Bayesian nonparametric inference for the three-class Youden index and its associated optimal cut-points

**Citation for published version:**
Calhau Fernandes Inacio De Carvalho, V & Branscum, AJ 2018, 'Bayesian nonparametric inference for the three-class Youden index and its associated optimal cut-points' Statistical Methods in Medical Research, vol 27, no. 3, pp. 689-700. DOI: 10.1177/0962280217742538

**Digital Object Identifier (DOI):**
10.1177/0962280217742538

**Link:**
Link to publication record in Edinburgh Research Explorer

**Document Version:**
Peer reviewed version

**Published In:**
Statistical Methods in Medical Research

Journal Title
XX(X):1–13
©The Author(s) 0000
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

# Bayesian nonparametric inference for the three-class Youden index and its associated optimal cut-points

Vanda Inácio de Carvalho[1] and Adam J. Branscum[2]

**Abstract**

The three-class Youden index serves both as a measure of medical test accuracy and a criterion to choose the optimal pair of cutoff values for classifying subjects into three ordinal disease categories (e.g., no disease, mild disease, advanced disease). We present a Bayesian nonparametric approach for estimating the three-class Youden index and its corresponding optimal cutoff values based on Dirichlet process mixtures, which are robust models that can handle intricate features of distributions for complex data. Results from a simulation study are presented and an application to data from the Trail Making Test to assess cognitive impairment in Parkinson's disease patients is detailed.

**Keywords**

Diagnostic test, Dirichlet process mixtures, optimal cutoff, true class fraction, Youden index

## 1 Introduction

Accurate diagnosis of disease is of fundamental importance in clinical practice and medical research. The ability of a screening or diagnostic test to distinguish between different stages or conditions (e.g., disease versus non-disease) should be rigorously evaluated before a test is routinely used in practice. The receiver operating characteristic (ROC) curve is a popular tool for evaluating the discriminatory ability of continuous tests. Two commonly used summary indices of medical test accuracy are the area under the ROC curve (AUC) and the Youden index (Youden 1950). A useful advantage of the Youden index over the AUC is that, besides serving as a summary measure of test accuracy, it provides a criterion for choosing the optimal cutoff to screen subjects in practice. A vast amount of research has been devoted to the ROC curve, AUC, and the Youden index for the case of two diagnostic outcome categories (for an overview see, for instance, Pepe 1997 and Zhou et al. 2011).

Disease progression is a dynamic process. In clinical practice, physicians often face situations that require decisions among three (or more) diagnostic alternatives. One or more intermediate transitional stages might exist prior to full disease onset, as is the case for many neurological disorders. For instance, Nakas et al. (2004) used the average score on 8 neuropsychological tests as a marker of the presence of HIV-related cognitive dysfunction to discriminate between HIV patients with clinical symptoms of cognitive dementia, patients exhibiting minor neurological symptoms, and neurologically unimpaired patients. Xiong et al. (2006) used a battery of psychometric tests to distinguish between subjects with cognitive changes due to normal aging, subjects with mild cognitive impairment or early stage Alzheimer's disease, and subjects with severe dementia caused by Alzheimer's disease. In this paper we are concerned with the discriminatory ability of the Trail Making Test (a visual search test that has been extensively used since the 1950s for neuropsychological assessment) to discriminate between Parkinson's disease subjects who present normal cognition, mild cognitive impairment, and dementia/severe impairment.

[1]School of Mathematics, University of Edinburgh, Scotland, UK
[2]College of Public Health and Human Sciences, Oregon State University, USA

**Corresponding author:**
Vanda Inácio de Carvalho, School of Mathematics, University of Edinburgh, EH9 3FD, UK.
Email: Vanda.Inacio@ed.ac.uk

The focus of this paper is the three-class Youden index and its associated pair of optimal cutoff values, which can be used to place subjects into one of three disease classes. Parametric and empirical estimators of the three-class Youden index were developed by Nakas et al. (2010). Luo and Xiong (2013) proposed a kernel-based estimator of the three-class Youden index. We use a Bayesian nonparametric approach to obtain data-driven inference for the three-class Youden index and its optimal cutoffs, based on a flexible Dirichlet process mixture (DPM) model. DPMs are robust models that can adapt to intricate distributional features, such as multimodality, skewness, and/or extreme variability, without the need to know of their existence in advance. Hence, the DPM model we present is a widely applicable approach to inference for the three-class Youden index that can be used for many populations and for a large number of diseases and continuous diagnostic measures. Markov chain Monte Carlo (MCMC) simulation is used to generate samples from posterior distributions and statistical inference does not depend on asymptotic theory or the bootstrap. Recent developments of Bayesian nonparametric models that have been successfully applied in medical diagnostic research abound (e.g., Erkanli et al. 2006; Branscum et al. 2008; Gu et al. 2008; Inácio et al. 2011; Inácio de Carvalho et al. 2013; Rodríguez and Martínez 2014; Branscum et al. 2015; Hwang and Chen 2015; Inácio de Carvalho et al. 2017).

The remainder of the paper is organised as follows. In the next section we introduce preliminary concepts and terminology. Section 3 presents our novel approach to estimating the three-class Youden index based on DPM models. The performance of our method is assessed in Section 4 using simulated data, and Section 5 applies our approach to data from a Trail Making Test of cognitive impairment in Parkinson's disease patients. Concluding remarks are provided in Section 6.

## 2 Preliminaries

We assume that there exists three ordered diagnostic groups (e.g., no disease, mild disease, advanced disease) and that each subject in the population belongs to one of the groups. A diagnostic test with continuous-scale outcomes is used for classification. Without loss of generality, we assume that subjects from group 3 tend to have higher test outcomes than subjects in group 2 who tend to have higher test outcomes than group 1 subjects. Let $Y_1$, $Y_2$, and $Y_3$ be continuous random variables denoting test outcomes in groups 1, 2, and 3, respectively, with $F_1$, $F_2$, and $F_3$ being the corresponding cumulative distribution functions. For any pair of ordered thresholds $(c_1, c_2)$, with $c_1 < c_2$, the probabilities of correct classification into each group are given by

$$\begin{aligned}
p_1(c_1, c_2) &= \Pr(Y_1 \leq c_1) = F_1(c_1),\\
p_2(c_1, c_2) &= \Pr(c_1 < Y_2 \leq c_2) = F_2(c_2) - F_2(c_1), \text{ and}\\
p_3(c_1, c_2) &= \Pr(Y_3 > c_2) = 1 - F_3(c_2).
\end{aligned}$$

The three-class Youden index (Nakas et al. 2010; Luo and Xiong 2013) is

$$\begin{aligned}
\mathrm{YI}_3 &= \max_{c_1 < c_2} \{p_1(c_1, c_2) + p_2(c_1, c_2) + p_3(c_1, c_2) - 1\}\\
&= \max_{c_1 < c_2} \{F_1(c_1) + F_2(c_2) - F_2(c_1) - F_3(c_2)\}. \qquad (1)
\end{aligned}$$

The pair of cutoff values that correspond to the Youden index,

$$(c_1^*, c_2^*) = \arg\max\{F_1(c_1) + F_2(c_2) - F_2(c_1) - F_3(c_2)\},$$

is considered optimal and can be used to aid classifying subjects in practice. The three-class Youden index, as defined in (1), falls in the range of $[0, 2]$. For a useless diagnostic test in which the three distributions completely overlap, $p_1(c_1, c_2) + p_2(c_1, c_2) + p_3(c_1, c_2) = 1$ and thus, $\mathrm{YI}_3 = 0$. The other extreme case involves a perfect classifier in which the three distributions are completely separated and $\mathrm{YI}_3 = 2$. Values between 0 and 2 correspond to different degrees of stochastic ordering between $F_1$, $F_2$, and $F_3$ (see Figure 1 of the Supplementary Material); the closer $\mathrm{YI}_3$ is to 2, the better the classification accuracy.

As noted by Nakas et al. (2010), $\mathrm{YI}_3$ can be interpreted as the maximum overall correct classification level when equal weight is given to the three correct classification probabilities. In general, unequal weights can be used to reflect the relative importance of the three classes, yielding

$$\mathrm{YI}_3^{\omega} = \max_{c_1 < c_2} \{\omega_1 p_1(c_1, c_2) + \omega_2 p_2(c_1, c_2) + \omega_3 p_3(c_1, c_2)\},$$

where $\omega_1$, $\omega_2$, and $\omega_3$ are weights associated with the classification probabilities. The focus of the rest of the paper is on $\mathrm{YI}_3$, but the methods presented easily extend to $\mathrm{YI}_3^{\omega}$.

## 3 Methods

From equation (1), it follows that an accurate estimate of $\mathrm{YI}_3$ can be obtained by accurately estimating the distribution functions of test outcomes for each group. Therefore, the models used for $F_1$, $F_2$, and $F_3$ must be flexible enough to encompass a wide range of biomarker distributions. In general, models based on Dirichlet process mixtures of normal distributions are successful at flexibly representing distribution functions of complex data. To motivate the Bayesian nonparametric mixture model, we start with a formulation based on finite mixtures of normal distributions, which are known to approximate any continuous distribution (Lo 1984).

Let $(y_{11},\dots,y_{1n_1})$, $(y_{21},\dots,y_{2n_2})$, and $(y_{31},\dots,y_{3n_3})$ be independent (within and between groups) samples of size $n_1$, $n_2$, and $n_3$ from groups 1, 2, and 3, respectively, with

$$y_{11},\dots,y_{1n_1} \mid F_1 \overset{\text{iid}}{\sim} F_1, \quad y_{21},\dots,y_{2n_2} \mid F_2 \overset{\text{iid}}{\sim} F_2, \quad y_{31},\dots,y_{3n_3} \mid F_3 \overset{\text{iid}}{\sim} F_3.$$

A finite normal mixture model posits

$$F_d(\cdot) = \sum_{k=1}^{K} p_{dk}\Phi(\cdot \mid \mu_{dk},\sigma^2_{dk}) \quad \text{for} \quad d \in \{1,2,3\}, \tag{2}$$

where $\Phi(z \mid \mu,\sigma^2)$ denotes the cumulative distribution function of the normal distribution with mean $\mu$ and variance $\sigma^2$ evaluated at $z$. Under this framework, each continuous test outcome arises from one of $K$ mixture components, which each have a specific mean and variance. For ease of presentation, we assume the number of components $K$ is the same across groups. The model in (2) can be equivalently written as

$$F_d(\cdot) = \int \Phi(\cdot \mid \mu,\sigma^2)\mathrm{d}G_d(\mu,\sigma^2),$$

where $G_d$ is a discrete (mixing) distribution given by

$$G_d(\cdot) = \sum_{k=1}^{K} p_{dk}\delta_{(\mu_{dk},\sigma^2_{dk})}(\cdot), \tag{3}$$

and where $\delta_a$ denotes a point mass at $a$. To proceed with Bayesian inference, prior distributions are required for the weights, means, and variances. The vector of weights $(p_{d1},\dots,p_{dK})$ is often assigned a Dirichlet prior distribution and, leveraging conjugacy properties, the prior for $(\mu_{dk},\sigma^2_{dk})$, say $G^*_d(\mu,\sigma^2)$, often is a normal-inverse-gamma distribution. Placing a prior distribution on $\{(p_{dk},\mu_{dk},\sigma^2_{dk}) \mid k = 1,\dots,K\}$ is equivalent to placing a prior distribution on the mixing distribution $G_d$. Finite mixture modeling provides a very flexible framework for density and distribution function estimation. However, choosing the number of mixture components $K$ is not a trivial task in general. It is common to use the value that optimizes a selection criterion across candidate models with different numbers of components. Another approach is to place a prior on $K$, but this can be computationally difficult to implement efficiently in practice (e.g., involving reversible jump MCMC).

The powerful alternative we use involves a Bayesian nonparametric Dirichlet process (Ferguson 1973) prior for $G_d$, which induces a DPM of normal distributions on $F_d$. In addition to the theoretical and practical advantage of the Dirichlet process (DP) mixture of normal model having full support on the space of absolutely continuous distributions (Lo 1984), the DP prior has the practical advantage of automatically determining the number of components that best fits a given data set. We write $G_d \sim \mathrm{DP}(\alpha_d, G^*_d)$ to indicate that $G_d$ follows a DP prior. The prior mean $E(G_d) = G^*_d$ is a parametric base/centering distribution. The positive precision parameter $\alpha_d$ determines, among other important characteristics, the variation of $G_d$ around the prior mean $G^*_d$, with smaller (larger) values of $\alpha_d$ implying higher (lower) uncertainty. According to Sethuraman (1994), the DP prior can be represented as

$$G_d(\cdot) = \sum_{k=1}^{\infty} p_{dk}\delta_{(\mu_{dk},\sigma^2_{dk})}(\cdot), \tag{4}$$

where $(\mu_{dk},\sigma^2_{dk}) \overset{\text{iid}}{\sim} G^*_d$ and $p_{d1} = v_{d1}$, $p_{dk} = v_{dk}\prod_{m=1}^{k-1}(1 - v_{dm})$ for $k \geq 2$, with $v_{dk} \overset{\text{iid}}{\sim} \mathrm{Beta}(1,\alpha_d)$, independently of $\{(\mu_{dk},\sigma^2_{dk}) \mid k \geq 1\}$. Notice that under Sethuraman's representation, uncertainty about each weight parameter $p_{dk}$ is induced from uncertainty about the $v_{dk}$'s.

The proposed model for the distribution function of test outcomes in group $d$ is thus given by a DPM of normals, namely

$$F_d(\cdot) = \int \Phi(\cdot \mid \mu, \sigma^2)\mathrm{d}G_d(\mu, \sigma^2), \quad G_d \sim \mathrm{DP}(\alpha_d, G_d^*), \quad d \in \{1, 2, 3\}. \tag{5}$$

Due to conjugacy properties, we take $G_d^*(\mu, \sigma^2) = \mathrm{N}(\mu \mid a_{\mu_d}, b^2_{\mu_d})\mathrm{IG}(\sigma^2 \mid a_{\sigma^2_d}, b_{\sigma^2_d})$. For ease of posterior simulation and because it provides a highly accurate approximation, we used a truncated version of the DPM model (Ishwaran and James 2001). Specifically, the mixing distribution $G_d$ in (4) is replaced with

$$G_d^L(\cdot) = \sum_{k=1}^{L} p_{dk}\delta_{(\mu_{dk}, \sigma^2_{dk})}(\cdot),$$

with $L$ being pre-specified and where the $p_{dk}$'s result from a truncated version of the stick-breaking construction: $p_{d1} = v_{d1}$, $p_{dk} = v_{dk}\prod_{m=1}^{k-1}(1 - v_{dm})$ for $k = 2, \ldots, L$, $v_{d1}, \ldots, v_{d,L-1} \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \alpha_d)$, and $v_{dL} = 1$. An appropriate value of $L$ can be determined by considering the properties of the high order weight values in the infinite sum representation (4). For instance, $E\left(\sum_{k=L+1}^{\infty} p_{dk} \mid \alpha_d\right) = \alpha_d^L(1 + \alpha_d)^{-L}$ (Ishwaran and Zarepour 2000). Setting $\alpha_d = 1$ and $L = 20$, as in our simulation study and application, leads to $E(\sum_{k=L+1}^{\infty} p_{dk}) \doteq 0$. The truncated DPM model for test outcomes can be expressed as

$$F_d(\cdot) = \sum_{k=1}^{L} p_{dk}\Phi(\cdot \mid \mu_{dk}, \sigma^2_{dk}), \tag{6}$$

where $\mu_{dk} \overset{\text{iid}}{\sim} \mathrm{N}(a_{\mu_d}, b^2_{\mu_d})$ and $\sigma^2_{dk} \overset{\text{iid}}{\sim} \mathrm{IG}(a_{\sigma^2_d}, b_{\sigma^2_d})$ for $k = 1, \ldots, L$, and the weights $p_{dk}$ arise from the truncated stick-breaking representation described above. Note that $L$ is not the exact number of components we expect to observe but instead an upper bound on the number of components.

We use configuration variables to identify the label of the mixture component to which the $i$th subject from group $d$ belongs. Let $S_{di} = k$ denote that the $i$th subject in group $d$ is allocated to component $k$ (for $d = 1, 2, 3$, $i = 1, \ldots, n_d$, and $k = 1, \ldots, L$). Then, the model can be written as

$$\begin{aligned}
y_{di} \mid \boldsymbol{\mu}_d, \boldsymbol{\sigma}^2_d, \mathbf{S}_d &\overset{\text{ind.}}{\sim} \mathrm{N}(\mu_{d,S_{di}}, \sigma^2_{d,S_{di}}), \quad i = 1, \ldots, n_d, \quad d \in \{1, 2, 3\},\\
\Pr(S_{di} = k \mid \mathbf{v}_d) &= p_{dk}, \quad k = 1, \ldots, L,\\
(\mu_{dk}, \sigma^2_{dk}) &\overset{\text{iid}}{\sim} \mathrm{N}(a_{\mu_d}, b^2_{\mu_d})\mathrm{IG}(a_{\sigma^2_d}, b_{\sigma^2_d}),
\end{aligned}$$

where $\boldsymbol{\mu}_d = (\mu_{d1}, \ldots, \mu_{dL})$, $\boldsymbol{\sigma}^2_d = (\sigma^2_{d1}, \ldots, \sigma^2_{dL})$, $\mathbf{S}_d = (S_{d1}, \ldots, S_{dn_d})$, and $\mathbf{v}_d = (v_{d1}, \ldots, v_{dL})$. The blocked Gibbs sampler was used to simulate draws from the posterior distribution; computational details are provided in the Appendix.

Posterior estimates of $\mathrm{YI}_3$ were obtained by using (1) and (6), and the pair of optimal cutoffs $(c_1^*, c_2^*)$ is the input that returns the maximum. A grid search was used to identify the maximum.

## 4 Simulation study

A simulation study was conducted to evaluate the performance of our nonparametric approach to estimating the three-class Youden index and its associated optimal cutoff values. For each of four scenarios, 100 data sets were generated using sample sizes of $(n_1, n_2, n_3) = (50, 50, 50)$, $(n_1, n_2, n_3) = (100, 100, 100)$, and $(n_1, n_2, n_3) = (200, 200, 200)$.

### *4.1 Simulation scenarios*

We considered the four scenarios listed in Table 1. Scenario 1 corresponds to the simple situation where test outcomes from the three groups follow normal distributions. In Scenario 2, data from the three groups follow different gamma distributions, while in Scenario 3, test outcomes arise from different distributional families. Scenario 4 considers the common setting of mixture distributions for test outcome data.

### 4.2 Models

Regarding prior information, $\alpha_d$ ($d \in \{1,2,3\}$) was set equal to one, which according to Hanson (2006) is the default value in the absence of prior information on the number of components needed to adequately describe $F_d$. To facilitate prior specification of the hyperparameters associated with the centering distribution, data were scaled by dividing by the standard deviation when fitting the model. We transformed back to the original scale when presenting the results. For the normal-inverse-gamma prior, we set $a_{\mu_d} = 0$, $b^2_{\mu_d} = 100$, $a_{\sigma^2_d} = 2$, and $b_{\sigma^2_d} = 0.5$. This configuration leads to relatively vague prior distributions for the $\mu_{dk}$'s and $\sigma^2_{dk}$'s. Note that the variance $b^2_{\mu_d}$ is large and that $a_{\sigma^2_d} = 2$ leads to a prior with infinite variance (hence, in some sense, vague) that is centered around a finite mean ($b_{\sigma^2_d} = 0.5$). The prior on $\sigma^2_{dk}$ therefore favours variances less than one; note that the scaled data has a marginal variance of one, so the within-component variance $\sigma^2_{dk}$ is expected to be smaller than the marginal variance. Finally, we capped the maximum number of components at $L = 20$ and, thus, a maximum of 20 normal distributions were used to approximate $F_d$.

As suggested by a referee, we included a comparison with the Bayesian bootstrap (BB)(Rubin 1981). Briefly, the BB corresponds to the limit of a Dirichlet process when the precision parameter $\alpha_d$ converges to zero (Gasparini 1995). The BB we implemented used 1000 resampled values. Additional details about the BB are given in Section 1 of the Supplementary Material.

We also compared the performance of the DPM estimator to a nonparametric frequentist kernel estimator and an empirical estimator. Specifically, for the kernel method, we used

$$\widehat{F}_d(c) = \frac{1}{n_d}\sum_{i=1}^{n_d}\Phi\left(\frac{c-y_{di}}{h_d}\right), \qquad d \in \{1,2,3\}.$$

We set $h_d = 0.9\min\{\text{SD}(\mathbf{y}_d), \text{IQR}(\mathbf{y}_d)/1.34\}n_d^{-0.2}$, where $\text{SD}(\mathbf{y}_d)$ and $\text{IQR}(\mathbf{y}_d)$ are the standard deviation and interquantile range, respectively, of $\mathbf{y}_d = (y_{d1}, \ldots, y_{dn_d})$. The empirical method estimates $F_d$ by its empirical distribution function. In addition, for Scenario 1 (where test outcomes in each group follow a normal distribution), we also included a comparison to a model involving independent parametric normal distributions in order to assess the efficiency of our nonparametric estimator in this context. The prior distributions used in the parametric normal model aligned with those used in the DPM model. Specifically, the normal model was

$$y_{di} \overset{\text{iid}}{\sim} \text{N}(\tilde{\mu}_d, \tilde{\sigma}^2_d), \quad \tilde{\mu}_d \sim \text{N}(a_{\tilde{\mu}_d}, b^2_{\tilde{\mu}_d}), \quad \tilde{\sigma}^2_d \sim \text{IG}(a_{\tilde{\sigma}^2_d}, b_{\tilde{\sigma}^2_d}), \quad i = 1, \ldots, n_d, \quad d \in \{1,2,3\},$$

with $a_{\tilde{\mu}_d} = 0$, $b^2_{\tilde{\mu}_d} = 100$, $a_{\tilde{\sigma}^2_d} = 2$, and $b_{\tilde{\sigma}^2_d} = 0.5$. For both the DPM and normal models, 1000 posterior samples were retained after a burn-in period of 500 iterations of the Gibbs sampler.

### 4.3 Results

The estimated Youden index and associated optimal cutoff values for Scenarios 1-4 are presented in Figures 1–4. Specifically, for each scenario and sample size considered, we present a boxplot of the estimates produced by each method (posterior means in the case of the DPM estimator) along with the true value. In Scenario 1, we can appreciate the minor loss in efficiency of our nonparametric estimator when the normal model holds (Figure 1), which is a small price to pay for the benefit of the robustness that leads to accurate data-driven estimates under increasingly complex scenarios (Figures 2–4). The DPM estimator outperformed the empirical estimator for most of the scenarios and sample sizes considered. The DPM estimator was on par with the kernel estimator, outperforming it in several cases; exceptions occurred in Scenario 3 for $c^*_1$ with a sample size of 200 and $c^*_2$ with a sample size of 50, and in Scenario 4 for $\text{YI}_3$ with a sample size of 50. The BB method accurately estimated most of the optimal cutoff values, but in many cases it had similar poor performance as the empirical method in terms of estimating the Youden index. Lastly, as expected, posterior uncertainty associated with the DPM estimator decreased as the sample size increased.

## 5 Application

The trail making test (TMT) is a multifactorial, neuropsychological assessment of motor speed, visual speed, executive functioning, and scanning, sequencing and cognitive ability. The TMT is commonly used as a diagnostic indicator of cognitive deficiency and brain impairment. The TMT comprises two parts. Part A involves drawing lines to connect circled numbers in an increasing sequence (1–2–3 etc). Part B involves drawing lines to connect circled numbers and letters in an alternating numerical and alphabetical sequence (1–A–2–B etc). The goal of the test is for the subject to finish both parts as quickly as possible. Completion times are used as the primary performance metrics.

We analysed completion time data from Part A of the TMT for 245 subjects with Parkinson's disease (Bantis et al. 2017). Based on a battery of cognitive tests used for the characterisation of cognitive impairment, 170 subjects were classified as unimpaired (U), 52 subjects were classified as having mild cognitive impairment (MCI), and 23 subjects were classified as having dementia (D). Parkinson's disease patients who have dementia were expected to have slower completion times than those with MCI, and patients with MCI were expected to have slower completion times than those with no cognitive impairment. In symbols, the anticipated ordering of completion times is U $<$ MCI $<$ D. Figure 2 of the Supplementary Material presents histograms and variable-width boxplots of the completion times for each group.

We estimated the three-class Youden index and optimal cutoff values using the same DPM model and prior information as in the simulation study. Posterior inference was based on estimates calculated from 3500 Gibbs sampler iterates after a burn-in of 1500 realisations was discarded. Completion times were scaled by dividing by the standard deviation to fit the model, but the results are presented on the original scale of the data.

Density estimates (posterior means and 95$\%$ pointwise probability bands) track the histograms for each group (Figure 5). The estimated Youden index (95$\%$ posterior interval) is $1.2$ $(1.0, 1.4)$, which suggests that Part A TMT completion times have reasonable discriminatory capacity to distinguish between U, MCI, and D patients. The estimated optimal cutoff values (in seconds) are 47 $(43, 50)$ and 81 $(73, 91)$. Based on this analysis, Parkinson's disease patients with Part A TMT completion times less than 47 seconds would be classified as unimpaired, patients with completion times between 47 and 81 seconds would be classified as having MCI, and patients with completion times greater than 81 seconds would be classified in the D group.

Estimates of the Youden index under the parametric normal model, BB, kernel, and empirical method were similar to the estimate of $\mathrm{YI}_3$ from the DPM model. However, estimates of the optimal cutoff values varied for these approaches. Estimates of $c_1^*$ and $c_2^*$ were 51 $(48, 52)$ and 90 $(84, 97)$ from the parametric normal model, 48 $(39, 53)$ and 74 $(62, 88)$ from the BB method, 48 $(42, 52)$ and 81 $(72, 89)$ from kernel density estimation, and 50 $(39, 53)$ and 70 $(62, 90)$ from the empirical method. The DPM model and kernel method gave similar estimates of $c_1^*$ and $c_2^*$. However, the normal model, BB, and empirical method gave different estimates of $c_2^*$ than the DPM model. Compared to the DPM, the percent difference in the point estimates of $c_2^*$ were 11$\%$ (normal), 9% (BB), and 14$\%$ (empirical). Note that the confidence intervals for the kernel and empirical estimates were computed using a nonparametric bootstrap method with 500 resampled values, whereas the BB estimates were based on 1000 replicates.

A sensitivity analysis to prior specification was conducted and results are presented and discussed in Section 2 of the Supplementary Material. In short, while inferences for $c_1^*$ and $\mathrm{YI}_3$ remained essentially unchanged, estimates of $c_2^*$ were slightly more sensitive to prior input in a predictable manner (higher prior variance led to higher posterior variance).

## 6 Conclusion

We applied a Bayesian nonparametric approach based on Dirichlet process mixtures to estimate the three-class Youden index and the corresponding pair of optimal cutoff values. Our simulation study illustrated the ability of our approach to produce accurate estimates for a variety of data-generating distributions. The DPM methodology can be readily extended to estimate the Youden index and associated cutoff values in the case where more than three disease categories exist. In addition, Attwood et al. (2014) presented alternative criteria to the Youden index for estimating optimal cutoff values and our approach can be directly applied to those criteria.

As an alternative to the 'pure' Bayesian bootstrap of Rubin (1981) that we used in Sections 4 and 5, note that equation (1) can be written as

$$\mathrm{YI}_3 = \max_{t_1 \geq t_2} \{R_3(t_2) - t_1 - R_2(t_2) + R_2(t_1)\},$$

where $t_1 = 1 - F_1(c_1)$, $t_2 = 1 - F_1(c_2)$, and $R_2$ and $R_3$ are ROC functions, namely

$$R_2(t) = 1 - F_2\{F_1^{-1}(1-t)\}, \qquad R_3(t) = 1 - F_3\{F_1^{-1}(1-t)\}.$$

Estimation can proceed according to the method presented in Gu et al. (2008). However, the results in the simulation study (not presented but available from the authors) did not show substantial improvement over the BB method used in this paper.

### Acknowledgements

Vanda Inácio de Carvalho acknowledges support from FCT–Fundação para a Ciência e a Tecnologia, Portugal, through the project UID/MAT/00006/2013.

### References

Attwood K, Tian L and Xiong C. Diagnostic thresholds with three ordinal groups. *Journal of Biopharmaceutical Statistics* 2014; **24**: 608–633.

Bantis LE, Nakas CT, Reiser B, Myall D and Dalrymple-Alford, JC. Construction of joint confidence regions for the optimal true class fractions of Receiver Operating Characteristic (ROC) surfaces and manifolds. *Statistical Methods in Medical Research* 2017; **26**: 1429–1442.

Branscum AJ, Johnson WO, Hanson TE and Gardner IA. Bayesian semiparametric ROC curve estimation and disease diagnosis. *Statistics in Medicine* 2008; **1**: 2474–2496.

Branscum AJ, Johnson WO, Hanson TE, and Baron AT. Flexible regression models for ROC and risk analysis, with or without a gold standard. *Statistics in Medicine* 2015; **30**: 3997–4015.

Erkanli E, Sung M, Costello EJ and Angold A. Bayesian semi-parametric ROC analysis. *Statistics in Medicine* 2006; **25**: 3905–3928.

Ferguson TS. A Bayesian analysis of some nonparametric problems. *Annals of Statistics* 1973; **1**: 209–230.

Gasparini M. Exact multivariate Bayesian bootstrap distributions of moments. *Annals of Statistics* 1995; **23**: 762–768.

Geisser S and Eddy WF. A predictive approach to model selection. *Journal of the American Statistical Association* 1979; **74**: 153–160.

Gu J, Ghosal S and Roy A. Bayesian bootstrap estimation of ROC curve. *Statistics in Medicine* 2008; **27**: 5407–5420.

Hanson TE. Modeling censored lifetime data using a mixture of gammas baseline. *Bayesian Analysis* 2006; **1**: 575–594.

Hwang BS and Chen Z. 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* 2015; **110**: 923–934.

Inácio V, Turkman AA, Nakas CT, and Alonzo TA. Nonparametric Bayesian estimation of the three-way receiver operating characteristic surface. *Biometrical Journal* 2011; **53**: 1011–1024.

Inácio de Carvalho V, Jara A, Hanson TE and de Carvalho, M. Bayesian nonparametric ROC regression modeling. *Bayesian Analysis* 2013; **8**: 623–646.

Inácio de Carvalho V, de Carvalho M and Branscum AJ. Nonparametric Bayesian covariate-adjusted estimation of the Youden index. *Biometrics* (in press) DOI: 10.1111/biom.12686.

Ishwaran I and James LF. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* 2001; **96**: 161–173.

Ishwaran H and Zarepour M. Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika* 2000; **87**: 371–390.

Lo AY. On a class of Bayesian nonparametric estimates: I. Density estimates. *Annals of Statistics* 1984; **12**: 351–357.

Luo J and Xiong C. Youden index and associated cut-points for three ordinal diagnostic groups. *Communications in Statistics - Simulation and Computation* 2013; **42**: 1213–1234.

Nakas CT and Yiannoutsos CT. Ordered multiple-class ROC analysis with continuous measurements. *Statistics in Medicine* 2004; **23**: 3437–3449.

Nakas CT, Alonzo TA and Yiannoutsos CT. Accuracy and cut-off point selection in three-class classification problems using a generalization of the Youden index. *Statistics in Medicine* 2010; **29**: 2946–2955.

Nakas CT, Dalrymple-Alford JC, Anderson TJ and Alonzo TA. Generalization of Youden index for multiple class classification problems applied to the assessment of externally validated cognition in Parkinson disease screening. *Statistics in Medicine* 2013; **32**: 995–1003.

Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford Statistical Science Series, 1997.

Rodríguez A and Martínez JC. Bayesian semi-parametric estimation of covariate-dependent ROC curves. *Biostatistics* 2014; **15**: 353–369.

Rubin DB. The Bayesian bootstrap. *Annals of Statistics* 1981; **9**: 130–134.

Sethuraman J. A constructive definition of Dirichlet priors. *Statistica Sinica* 1994; **4**: 639–650.

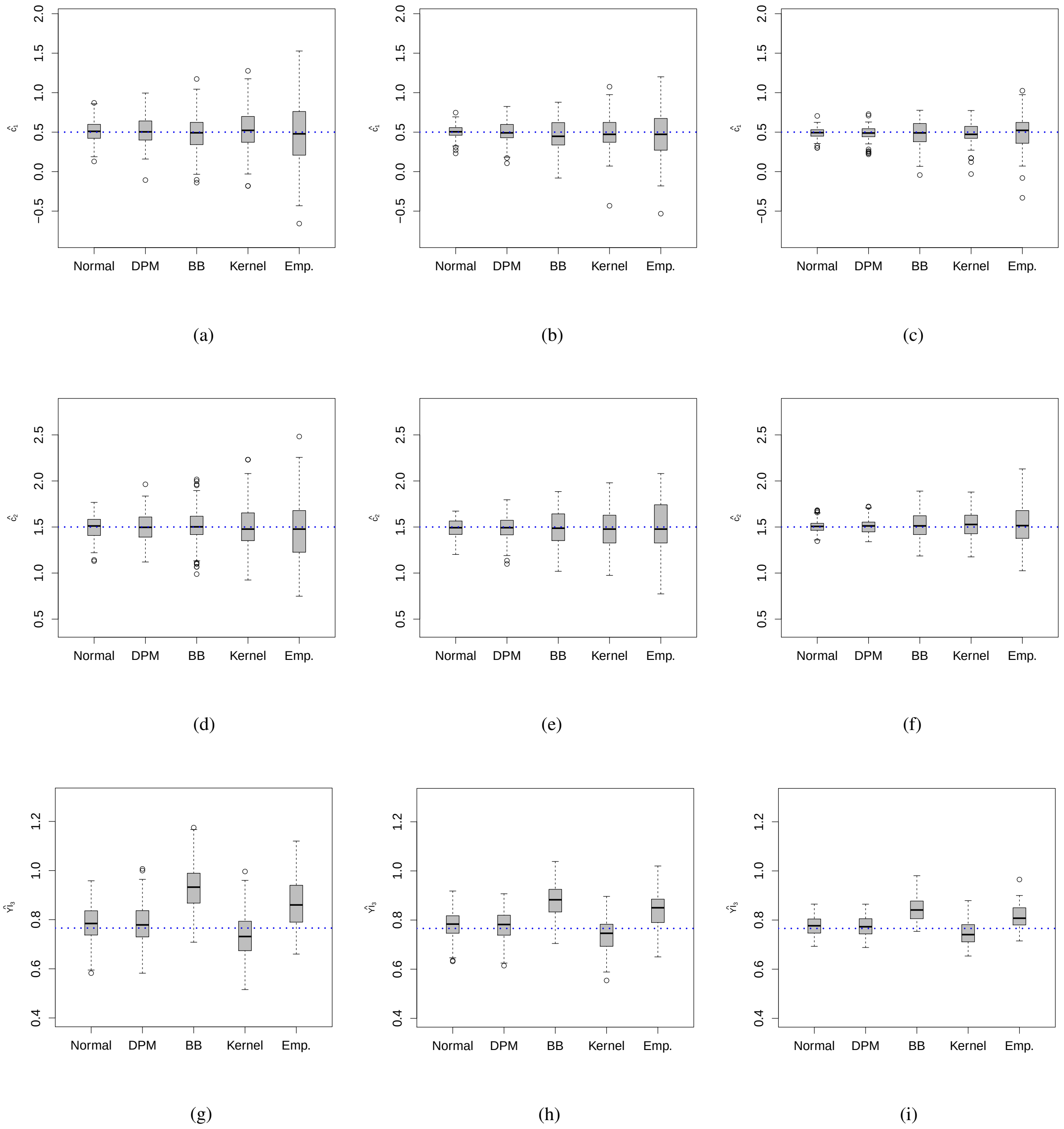
Xiong C, van Belle G, Miller JP and Morris JC. Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups. *Statistics in Medicine* 2006; **25**: 1251–1273.

Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; **3**: 32–35.

Zhou XH, Obuchowski NA and McClish DK. *Statistical Methods in Diagnostic Medicine*. Wiley Series in Probability and Statistics, 2nd edition, 2011.

| Scenario | $Y_1$ | $Y_2$ | $Y_3$ |
|---|---|---|---|
| 1 | $\mathrm{N}(0,1)$ | $\mathrm{N}(1,1)$ | $\mathrm{N}(2,1)$ |
| 2 | $\mathrm{Gamma}(2,1)$ | $\mathrm{Gamma}(3,1)$ | $\mathrm{Gamma}(5,2)$ |
| 3 | $t_2$ | $\mathrm{Beta}(2,2)$ | $\chi_1^2$ |
| 4 | $\frac{1}{2}\mathrm{N}(-1.5,0.5^2)+\frac{1}{2}\mathrm{N}(0.5,1)$ | $\frac{1}{2}\mathrm{N}(1,1)+\frac{1}{2}\mathrm{N}(4,1.5^2)$ | $\mathrm{N}(5,2^2)$ |

**Table 1.** Scenarios considered for the simulation study.

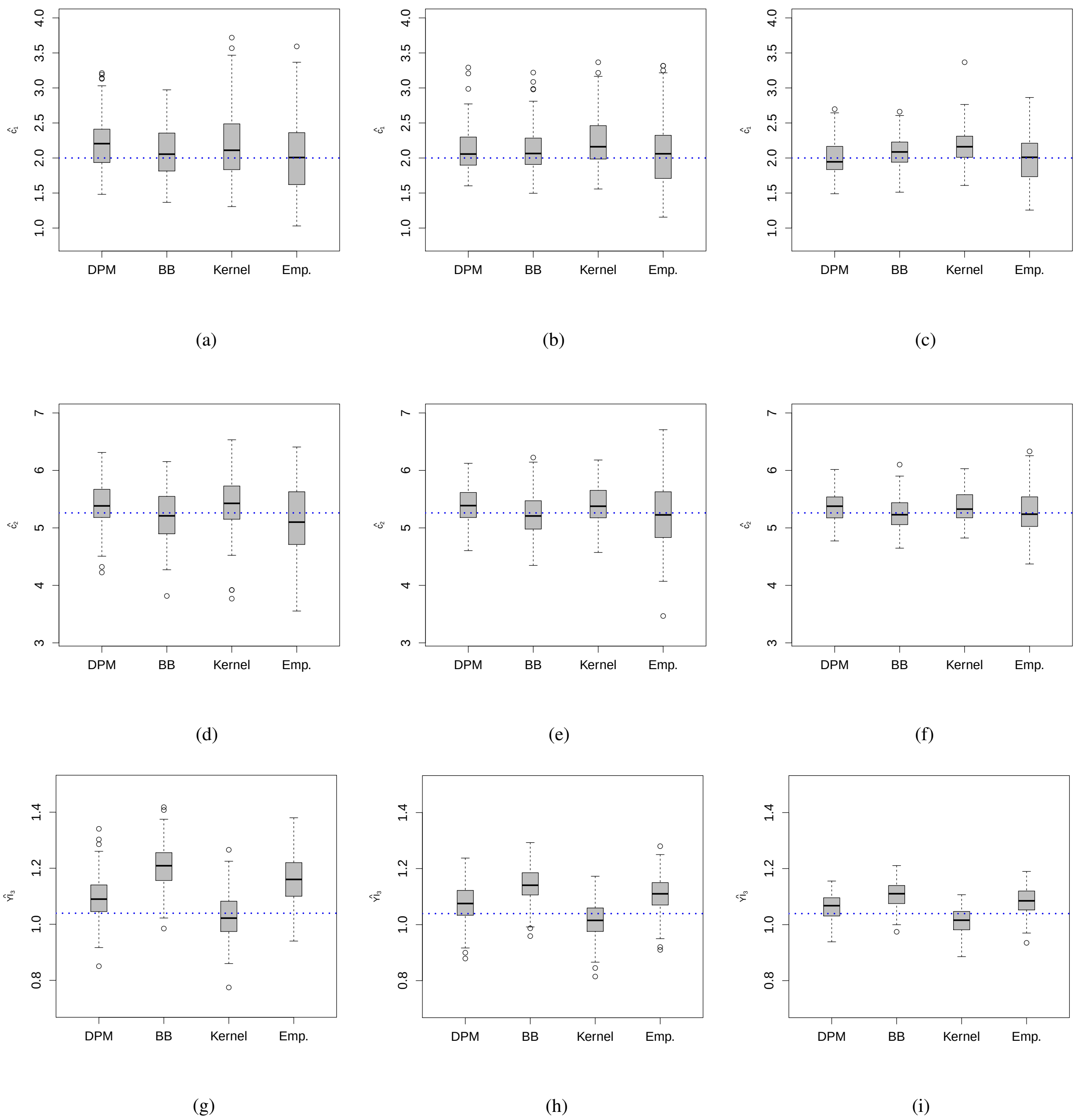


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

**Figure 1.** Boxplots summarising simulation results for estimates $\widehat{c}_1$ (row 1), $\widehat{c}_2$ (row 2), and $\widehat{\mathrm{YI}}_3$ (row 3) of the optimal cutoff values and Youden index in Scenario 1. The dotted blue line corresponds to the true value. Panels (a), (d), and (g): $n_1 = n_2 = n_3 = 50$. Panels (b), (e), and (h): $n_1 = n_2 = n_3 = 100$. Panels (c), (f), and (i): $n_1 = n_2 = n_3 = 200$. DPM=Dirichlet process mixture, BB=Bayesian bootstrap, Emp=Empirical estimator.

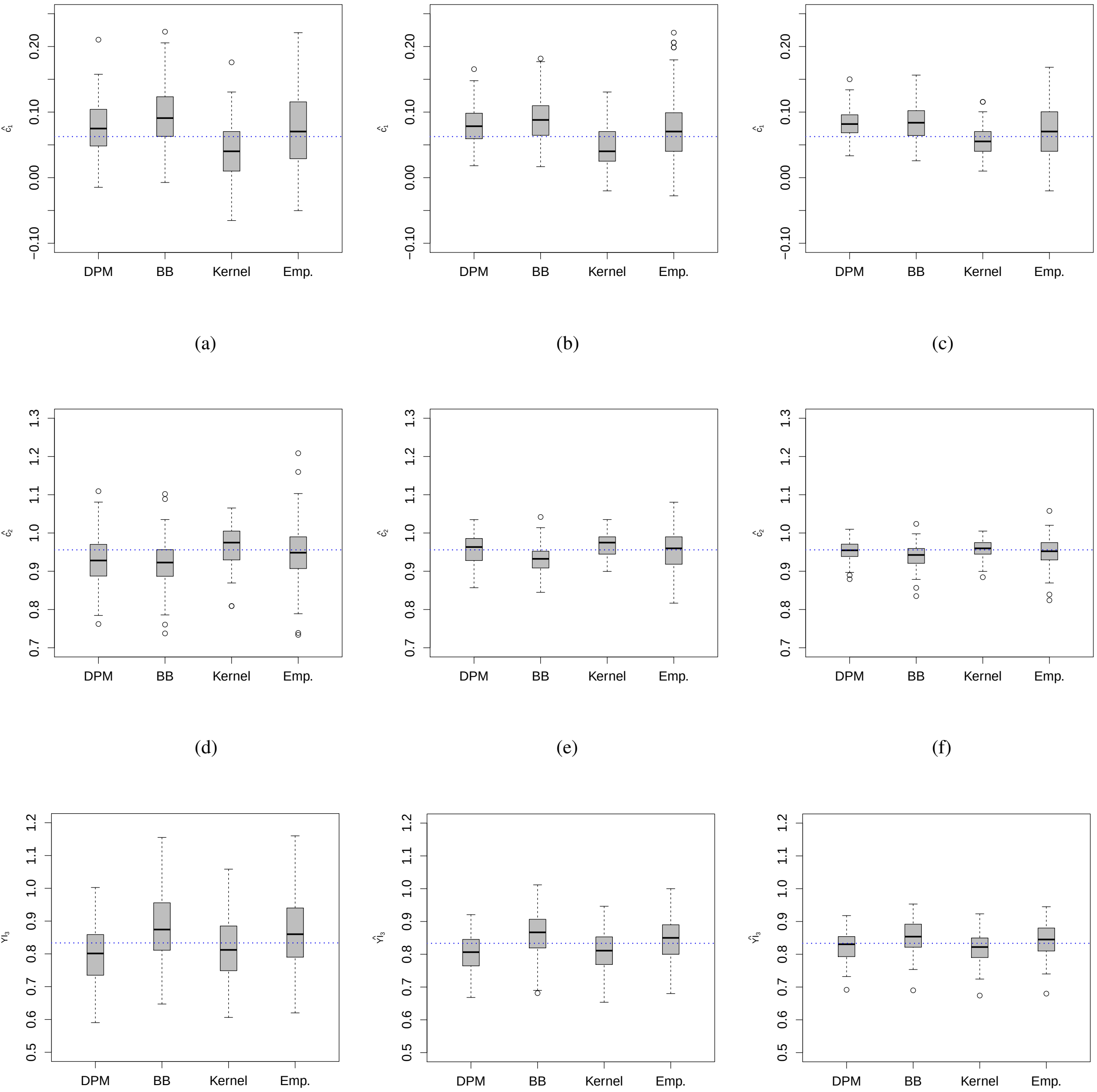


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

**Figure 2.** Boxplots summarising simulation results for estimates $\widehat{c}_1$ (row 1), $\widehat{c}_2$ (row 2), and $\widehat{\text{YI}}_3$ (row 3) of the optimal cutoff values and Youden index in Scenario 2. The dotted blue line corresponds to the true value. Panels (a), (d), and (g): $n_1 = n_2 = n_3 = 50$. Panels (b), (e), and (h): $n_1 = n_2 = n_3 = 100$. Panels (c), (f), and (i): $n_1 = n_2 = n_3 = 200$. DPM=Dirichlet process mixture, BB=Bayesian bootstrap, Emp=Empirical estimator.

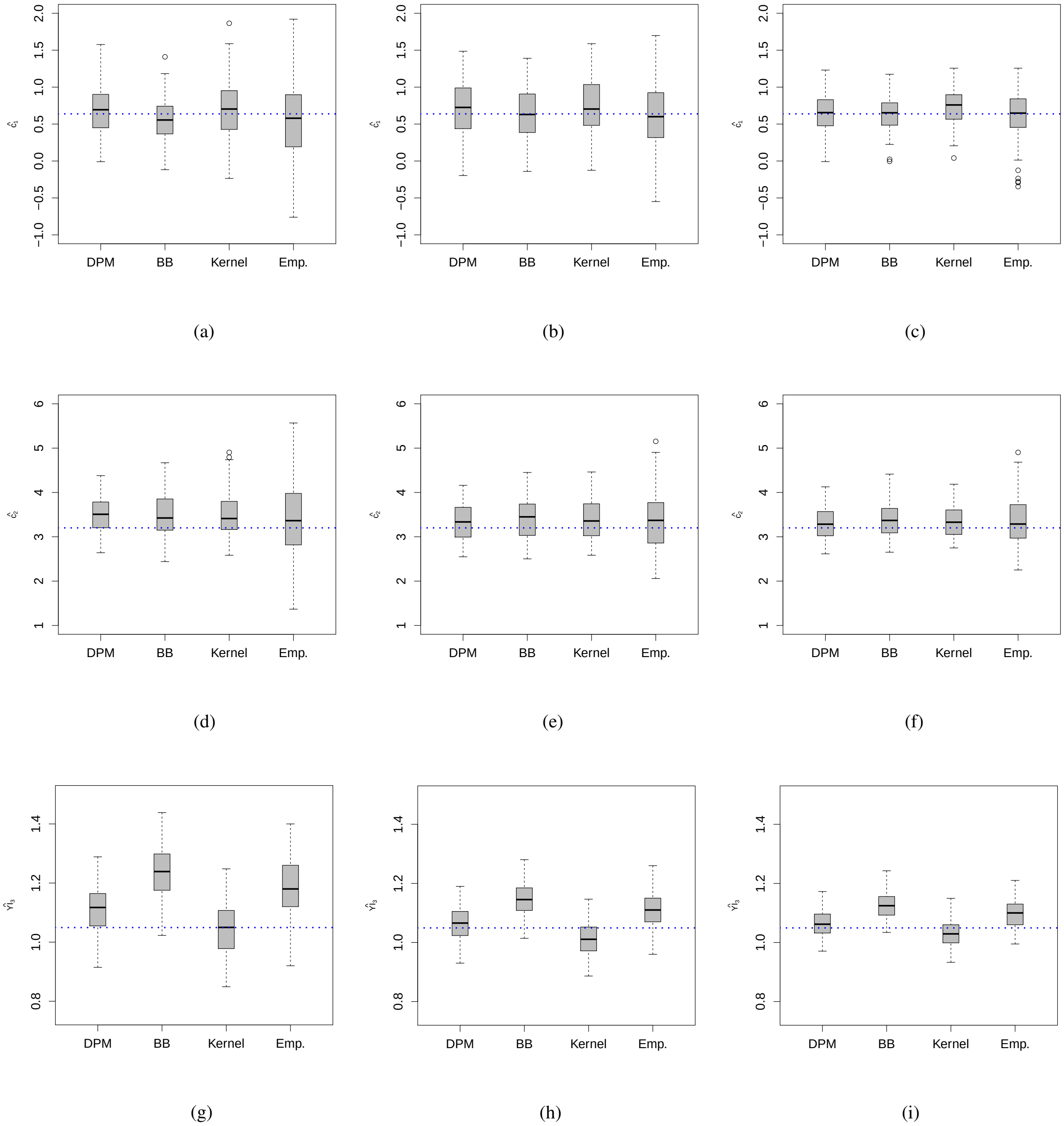


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

**Figure 3.** Boxplots summarising simulation results for estimates $\widehat{c}_1$ (row 1), $\widehat{c}_2$ (row 2), and $\widehat{\text{YI}}_3$ (row 3) of the optimal cutoff values and Youden index in Scenario 3. The dotted blue line corresponds to the true value. Panels (a), (d), and (g): $n_1 = n_2 = n_3 = 50$. Panels (b), (e), and (h): $n_1 = n_2 = n_3 = 100$. Panels (c), (f), and (i): $n_1 = n_2 = n_3 = 200$. DPM=Dirichlet process mixture, BB=Bayesian bootstrap, Emp=Empirical estimator.

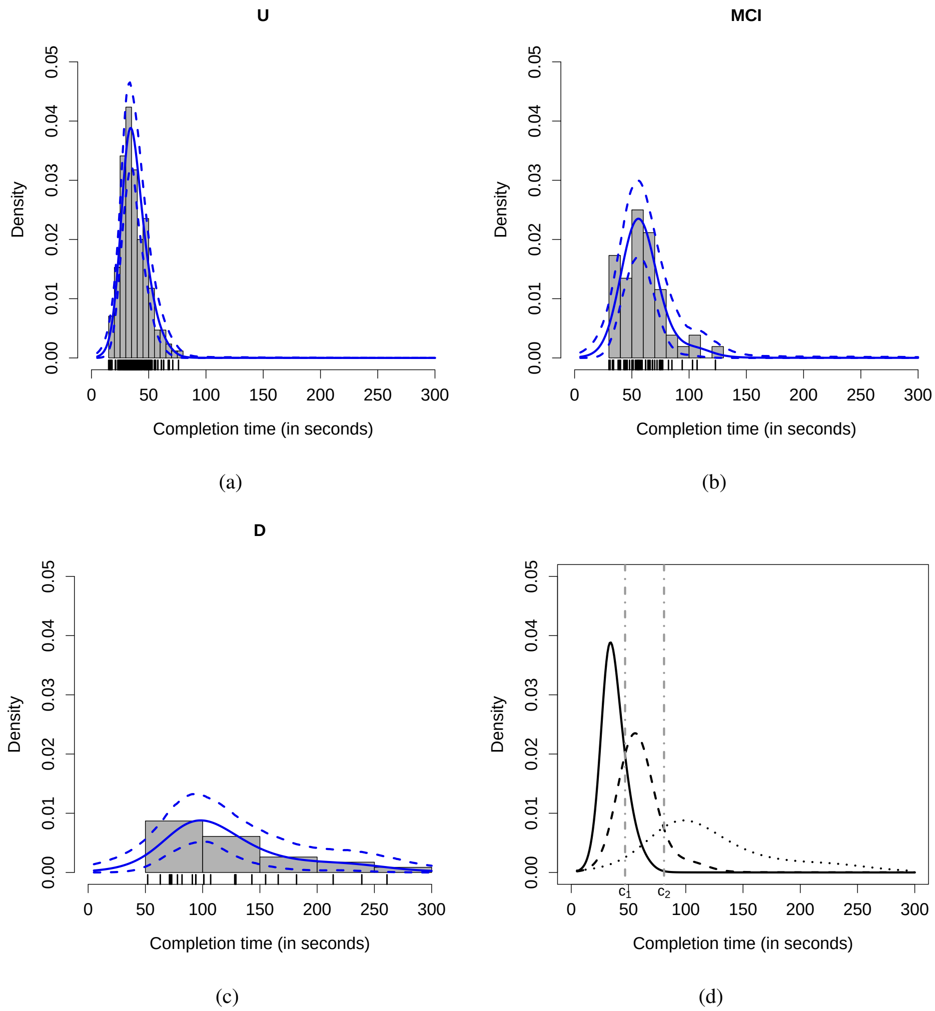


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

**Figure 4.** Boxplots summarising simulation results for estimates $\widehat{c}_1$ (row 1), $\widehat{c}_2$ (row 2), and $\widehat{\text{YI}}_3$ (row 3) of the optimal cutoff values and Youden index in Scenario 4. The dotted blue line corresponds to the true value. Panels (a), (d), and (g): $n_1 = n_2 = n_3 = 50$. Panels (b), (e), and (h): $n_1 = n_2 = n_3 = 100$. Panels (c), (f), and (i): $n_1 = n_2 = n_3 = 200$. DPM=Dirichlet process mixture, BB=Bayesian bootstrap, Emp=Empirical estimator.



(a)

(b)

(c)

(d)

**Figure 5.** Panels (a), (b), and (c): Density estimates (posterior means and $95\%$ pointwise posterior bands) from Dirichlet process mixture of normal modeling of trail making test completion times for Parkinson's disease patients who have dementia (D), mild cognitive impairment (MCI), or are cognitively unimpaired (U). Panel (d): Estimated densities (solid line for the U group, dashed line for the MCI group, and dotted line for the D group) along with the estimated optimal cutoff values.