ROC estimation and threshold selection criteria in three-class classification problems for clustered data

ROC estimation and threshold selection criteria in three-class classification problems for clustered data

Statistical Methods in Medical Research
XX(X):2–26

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Duc-Khanh To[1], Gianfranco Adimari[1], Monica Chiogna[2] and Davide Risso[1]

Abstract

Statistical evaluation of diagnostic tests, and, more generally, of biomarkers, is a constantly developing field, in which complexity of the assessment increases with complexity of the design under which data are collected. One particularly prevalent type of data is clustered data, where individual units are naturally nested into clusters. In these cases, bias can arise from omission, in the evaluation process, of cluster-level effects and/or individual covariates. Focussing on the three-class case and for continuous-valued diagnostic tests, we investigate how to exploit the clustered structure of data within a linear-mixed model approach, both when the assumption of normality holds and when it does not. We provide a method for estimation of covariate-specific ROC surfaces and discuss methods for the choice of optimal thresholds, proposing three possible estimators. A proof of consistency and asymptotic normality of the proposed threshold estimators is given. All considered methods are evaluated by extensive simulation experiments. As an application, we study the use of the *Lysosomal Associated Membrane Protein Family Member 5* (Lamp5) gene expression as biomarker to distinguish among three types of glutamatergic neurons.

Keywords

ROC analysis, clustered data, covariate adjustment, linear-mixed models, Box-Cox transformation

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

1 Introduction

Statistical evaluation of the accuracy of a diagnostic test (or biomarker) is an important step before its eventual wide-scale use. In the simplest setting of a biomedical study, the disease status typically has two classes, healthy and diseased, say, and for a such two-class setting, the receiver operating characteristic (ROC) curve, and indexes derived from it, have been widely used to evaluate the accuracy of a diagnostic test. See [1] and [2] for comprehensive reviews.

However, in many studies, the disease status, or, more generally, the condition to be classified, can have three classes, or even more. A prominent example is given by cancer studies, where the disease is usually staged in classes that identify the extent to which a cancer has developed. Contemporary practice foresees four cancer stages, with the fourth stage representing the most severe condition, but these can be grouped into fewer less-detailed stages in some circumstances (see [3] who analyse epithelial ovarian cancer data with a disease status classified as "benign", "early stage" or "late stage"). Recent advances in molecular biology have opened new research avenues in the field of classification of multiple statuses or biological varieties. In immunology, for example, the expression of cell surface proteins is routinely used to distinguish different white blood cell populations (see [4]). Neuroscientists are often requested to distinguish among the different cell types present in the mammalian brain [5] on the basis of gene biomarkers, typically measured using gene expression assays, such as single-cell RNA sequencing. In fact, some of the most urgent problems arising in the biosciences can be regarded as classification or decision problems using complex and often very extensive data. This makes rigorous evaluation of classification accuracy a crucial issue.

In a three-class setting, the ROC surface represents a natural generalization of the ROC curve and is commonly used to evaluate the ability of a diagnostic test to distinguish among three classes (or levels) of a disease. Let Y denote a diagnostic test result, often measured on a continuous scale, and let Y_1, Y_2, Y_3 be the test result for subjects in class 1, 2 and 3, respectively. Without loss of generality, we assume that higher values of test result are associated to higher severity of the disease, and the severity of the disease grows with the class (i.e., class 3 is the worst). Given a pair of thresholds (t_1, t_2), with $t_1 < t_2$ in the range of diagnostic test results, three true class fractions (TCFs) can be defined as

[1] Department of Statistical Sciences, University of Padova, Italy
[2] Department of Statistical Sciences "Paolo Fortunati", University of Bologna, Italy

Corresponding author:
Duc-Khanh To, Department of Statistical Sciences, University of Padova, Via C. Battisti, 241 - 35121 Padova, Italy.
Email: duckhanh.to@unipd.it

Prepared using sagej.cls

$\mathrm{TCF}_1(t_1) = \Pr(Y_1 \leq t_1)$, $\mathrm{TCF}_2(t_1, t_2) = \Pr(t_1 < Y_2 \leq t_2)$ and $\mathrm{TCF}_3(t_2) = \Pr(Y_3 > t_2)$. Then, the ROC surface for the diagnostic test Y is obtained by plotting ($\mathrm{TCF}_1(t_1)$, $\mathrm{TCF}_2(t_1, t_2)$ and $\mathrm{TCF}_3(t_2)$) in a unit cube over all possible values of t_1 and t_2 [6]. The volume under the ROC surface (VUS) is usually considered as a summary measure of the diagnostic accuracy of the test.

Clearly, from an operational perspective, the choice of optimal thresholds is crucial, and various proposals have been formulated in this respect. Recently, [8] proposed an approach based on a generalization of the Youden index, GYI hereafter, in which the optimal pair of thresholds is chosen to maximize the sum of three TCFs (or total of correct classification rates). Alternatively, [10] proposed two selection criteria, named closest to perfection (CtP) criterion and max volume (MV) criterion. In the CtP approach, the optimal pair of thresholds is obtained by minimizing the distance, in the unit cube, between the point $(\mathrm{TCF}_1(t_1), \mathrm{TCF}_2(t_1, t_2), \mathrm{TCF}_3(t_2))$ and the corner $(1, 1, 1)$, which corresponds to perfect discrimination. The MV approach searches for thresholds t_1 and t_2 that maximize the volume of a box under the ROC surface, and the volume is defined as the product of the three TCFs. Other proposed approaches (e.g., adjusted Youden index, maximum determinant) can be found in [11].

Most of the known statistical methods for ROC analysis (estimation of the ROC surface, VUS and optimal pair of thresholds), consider a setting in which measurements on statistical units can be considered as realizations of independent random variables, and the diagnostic test is not influenced by any covariate. In several studies, however, statistical units are enrolled in clusters (e.g., families), and the diagnostic test can be affected by some covariates that characterize the units themselves. In such contexts, [12] used a linear mixed-effect model [13] (with normal assumption) to account for clusters and covariates effects, and proposed an approach to estimate the VUS. However, to the best of our knowledge, no methods are available to estimate ROC surface, nor selecting an optimal pair of thresholds or constructing confidence regions in a clustered-data setting.

In this paper, we discuss covariate-specific estimation of a ROC surface of a continuous diagnostic test with clustered data, and adapt to the clustered-data case the criteria based on GYI, CtP and MV approaches, in order to properly address the problem of selecting an optimal pair of thresholds. Our approach also allows to properly build confidence regions for the true class fractions $(\mathrm{TCF}_1(t_1), \mathrm{TCF}_2(t_1, t_2), \mathrm{TCF}_3(t_2))$ and for the optimal pair of thresholds. We employ the model in [12] under normality assumptions for the cluster effects and the error terms; then, we estimate the covariate-specific TCFs, the ROC surface, and the optimal pair of thresholds based on the modified GYI, CtP and MV methods. We discuss the asymptotic behavior of the proposed estimators and use asymptotic results to construct confidence regions. In order to relax the normality assumption about the distributions of the test results Y_1, Y_2 and Y_3, we resort to the class of Box-Cox transformations for the linear-mixed effect model [14]. For this situation, we also consider a bootstrap procedure to estimate the covariance matrix of the estimator of the optimal thresholds.

The performances of our proposed estimators are verified through several simulation experiments. An application to real data is also presented. Specifically, we reanalyze a subset of the data used by [15], as processed by the authors*. In the aforementioned study, the authors focussed on the visual and motor cortex of the mouse brain. Here, we restrict our attention to the classification of three types of glutamatergic neurons, namely Layer 2/3 Intratelencephalic (L2/3 IT), Layer 4 (L4) and Layer 5 Pyramidal Tract (L5 PT) neurons, using the *Lysosomal Associated Membrane Protein Family Member 5* (Lamp5) gene expression as biomarker.

The paper is organized as follows. In Section 2, we present the model settings and discuss model estimation. Methods proposed for estimating the TCFs, the ROC surface and for selecting the optimal pair of thresholds are presented in Section 3. The simulation study is described in Section 4 and the application is described in Section 5. Concluding remarks are left to Section 6.

2 Linear mixed-effect model

Suppose we are interested in evaluating the accuracy of a diagnostic test which is potentially useful to classify the cases into three categories, of a disease status, say. Let Y be the diagnostic test result, on a continuous scale, and let Y_1, Y_2, Y_3 be the test result for subjects in class 1, 2 and 3, respectively. Suppose to have p covariates, $X_1, \ldots, X_p$, say, possibly associated with the test Y.

Let c be the total number of clusters (for instance, families), randomly selected from the population. For the k-th cluster, $k = 1, \ldots, c$, let n_{ki} be the total number of subjects belonging to class i, $i = 1, 2, 3$ and let $n_k = n_{k1} + n_{k2} + n_{k3}$ be the total sample size within the cluster. Note that n_{ki} might be equal to 0 for some clusters. Clearly, we expect that measures in the same cluster may be dependent. In order to account for the clustering effect on the test result Y, as well as for covariates' effects, we consider the following linear mixed-effect model (see also [12]):

$$\begin{aligned} Y_1 &= \alpha_{k_1} + z_1^\top \boldsymbol{\beta}_1 + \varepsilon_1, \\ Y_2 &= \alpha_{k_2} + z_2^\top \boldsymbol{\beta}_2 + \varepsilon_2, \\ Y_3 &= \alpha_{k_3} + z_3^\top \boldsymbol{\beta}_3 + \varepsilon_3, \end{aligned} \tag{2.1}$$

where (Y_1, Y_2, Y_3) is a triplet of test scores from three randomly sampled subjects from the three disease classes, (k_1, k_2, k_3), $k_i \in \{1, \ldots, c\}$, are cluster memberships indicating the clusters from which Y_1, Y_2, Y_3 are observed, $z_i = (1, x_{1i}, \ldots, x_{pi})^\top$ are fixed (i.e., not random) covariates values, and

*the data are publicly available at https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-v1-and-alm-smart-seq

$\boldsymbol{\beta}_i = (\beta_{0i}, \beta_{1i}, \ldots, \beta_{pi})^\top$, $i = 1, 2, 3$, are vectors of parameters representing covariates effects. In model (2.1), α_k are random effects accounting for the presence of clusters, and ε_i are subject-level random errors. We assume that: (i) the random effects α_k and the subject-level random errors ε_i follow a normal distribution, i.e., $\alpha_k \sim \mathcal{N}(0, \sigma_c^2)$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ with $i = 1, 2, 3$; (ii) $\alpha_1, \alpha_2, \ldots, \alpha_c$ and $\varepsilon_1, \varepsilon_2, \varepsilon_3$ are all independent. These assumptions are standard in the linear mixed-effect modelling framework [13].

Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\beta}_3^\top)^\top$ with $\boldsymbol{\beta}_i = (\beta_{0i}, \beta_{1i}, \ldots, \beta_{pi})^\top$, and $\boldsymbol{\theta} = (\sigma_c, \sigma_1, \sigma_2, \sigma_3)^\top$ be the unknown parameters in model (2.1). In order to obtain an estimator $\widehat{\boldsymbol{\gamma}} = (\widehat{\boldsymbol{\beta}}^\top, \widehat{\boldsymbol{\theta}}^\top)^\top$ of $\boldsymbol{\gamma} = (\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top)^\top$, a restricted (or residual) maximum likelihood (REML) is frequently adopted [13]. In particular, we can write the restricted log-likelihood for the model (2.1) as

$$\begin{aligned} \ell_R(\boldsymbol{\gamma}) = \ell_R(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta}) \quad = \quad & -\frac{1}{2}\sum_{k=1}^{c}\left(\mathbf{Y}_k - \mathbf{Z}_k\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\right)^\top \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{Y}_k - \mathbf{Z}_k\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\right) - \frac{1}{2}\sum_{k=1}^{c}\log|\boldsymbol{\Sigma}_k| \\ & - \frac{1}{2}\log\left|\sum_{k=1}^{c}\mathbf{Z}_k^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{Z}_k\right|, \end{aligned} \tag{2.2}$$

where $\mathbf{Y}_k$ is the n_k-vector of test results within the k-th cluster, $\mathbf{Z}_k$ is $n_k \times 3(p+1)$ design matrix for the fixed effects within the k-th cluster, $\boldsymbol{\Sigma}_k = \sigma_c^2\mathbf{V}_k\mathbf{V}_k^\top + diag\{\sigma_1^2, \ldots, \sigma_1^2; \sigma_2^2, \ldots, \sigma_2^2; \sigma_3^2, \ldots, \sigma_3^2\}_{n_k}$ with $\mathbf{V}_k$ as $\mathbf{1}_{n_k}$, and

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \left(\sum_{k=1}^{c}\mathbf{Z}_k^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{Z}_k\right)^{-1}\sum_{k=1}^{c}\mathbf{Z}_k^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{Y}_k.$$

Maximizing the restricted log-likelihood function $\ell_R(\boldsymbol{\gamma})$ (2.2), gives the REML estimator $\widehat{\boldsymbol{\theta}}$ of the variance components vector $\boldsymbol{\theta}$. Then $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}})$. Theoretical results on consistency and asymptotic normality of the resulting estimator $\widehat{\boldsymbol{\gamma}} = (\widehat{\boldsymbol{\beta}}^\top, \widehat{\boldsymbol{\theta}}^\top)^\top$ are given in [13]: under some regularity conditions, the REML estimator $\widehat{\boldsymbol{\gamma}}$ asymptotically follows a normal distribution with mean $\boldsymbol{\gamma}$ and covariance matrix $\boldsymbol{\Lambda}$, i.e., $\widehat{\boldsymbol{\gamma}} \dot{\sim} \mathcal{N}(\boldsymbol{\gamma}, \boldsymbol{\Lambda})$. The asymptotic covariance matrix $\boldsymbol{\Lambda}$ can be consistently estimated by using the sandwich formula [16–18], i.e.,

$$\widehat{\boldsymbol{\Lambda}} = c^{-1}J^{-1}(\boldsymbol{\gamma})I(\boldsymbol{\gamma})J^{-1}(\boldsymbol{\gamma})\Big|_{\boldsymbol{\gamma}=\widehat{\boldsymbol{\gamma}}}, \tag{2.3}$$

where

$$J(\boldsymbol{\gamma}) = c^{-1}\mathbb{E}\left\{\frac{\partial^2\ell_{R,k}(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^\top}\right\}, \qquad I(\boldsymbol{\gamma}) = c^{-1}\mathbb{E}\left[\frac{\partial\ell_{R,k}(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}\left\{\frac{\partial\ell_{R,k}(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}\right\}^\top\right],$$

and $\ell_{R,k}(\boldsymbol{\gamma})$ is the k-th contribution to the restricted log-likelihood function $\ell_R(\boldsymbol{\gamma})$.

3 The proposal

3.1 Covariate-specific ROC surface estimation for clustered data

According to the model (2.1), at a given vector z of covariates values, the distribution of Y_i, $i = 1, 2, 3$, is normal with mean $z^\top \boldsymbol{\beta}_i$ and variance $\sigma_c^2 + \sigma_i^2$, i.e., $Y_i \sim \mathcal{N}(z^\top \boldsymbol{\beta}_i, \sigma_c^2 + \sigma_i^2)$ with $z = (1, x_1, \ldots, x_p)^\top$ and $i = 1, 2, 3$. We further assume that $z^\top \boldsymbol{\beta}_1 < z^\top \boldsymbol{\beta}_2 < z^\top \boldsymbol{\beta}_3$, i.e., that the stochastic dominance for the three classes holds at z. This is equivalent to the assumption that the covariate-specific VUS is greater than $1/6$.[6,7] It is worth noting that such assumption does not affect the theoretical developments that follow, thanks to consistency of the unconstrained estimator $\widehat{\boldsymbol{\gamma}}$, on which, for convenience, we will rely throughout the paper.

For given thresholds t_1 and t_2 ($t_1 < t_2$) and a vector z of covariates values, the covariate-specific TCFs are:

$$\begin{aligned} \mathrm{TCF}_1(t_1; z) &= \Phi\left(\frac{t_1 - z^\top \boldsymbol{\beta}_1}{\sqrt{\sigma_c^2 + \sigma_1^2}}\right), \\ \mathrm{TCF}_2(t_1, t_2; z) &= \Phi\left(\frac{t_2 - z^\top \boldsymbol{\beta}_2}{\sqrt{\sigma_c^2 + \sigma_2^2}}\right) - \Phi\left(\frac{t_1 - z^\top \boldsymbol{\beta}_2}{\sqrt{\sigma_c^2 + \sigma_2^2}}\right), \\ \mathrm{TCF}_3(t_2; z) &= 1 - \Phi\left(\frac{t_2 - z^\top \boldsymbol{\beta}_3}{\sqrt{\sigma_c^2 + \sigma_3^2}}\right), \end{aligned} \tag{3.1}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. The plot of $(\mathrm{TCF}_1(t_1; z), \mathrm{TCF}_2(t_1, t_2; z), \mathrm{TCF}_3(t_2; z))$, by varying the pair of thresholds (t_1, t_2), produces a covariate-specific ROC surface of Y at given vector z. Alternatively, by writing $\mathrm{TCF}_1(t_1; z) = p_1$ and $\mathrm{TCF}_3(t_2; z) = p_3$, the covariate-specific ROC surface can be defined as a function of (p_1, p_3), i.e.,

$$\begin{aligned} \mathrm{ROCs}(p_1, p_3; z) &= \Phi\left(\frac{\Phi^{-1}(1 - p_3)\sqrt{\sigma_c^2 + \sigma_3^2} + z^\top \boldsymbol{\beta}_3 - z^\top \boldsymbol{\beta}_2}{\sqrt{\sigma_c^2 + \sigma_2^2}}\right) \\ &\quad - \Phi\left(\frac{\Phi^{-1}(p_1)\sqrt{\sigma_c^2 + \sigma_1^2} + z^\top \boldsymbol{\beta}_1 - z^\top \boldsymbol{\beta}_2}{\sqrt{\sigma_c^2 + \sigma_2^2}}\right), \end{aligned} \tag{3.2}$$

if $\Phi^{-1}(p_1) < \frac{\Phi^{-1}(1-p_3)\sqrt{\sigma_c^2+\sigma_3^2} + z^\top \boldsymbol{\beta}_3 - z^\top \boldsymbol{\beta}_1}{\sqrt{\sigma_c^2+\sigma_1^2}}$; otherwise, $\mathrm{ROCs}(p_1, p_3; z) = 0$. Given a pair of thresholds (t_1, t_2) and a pair (p_1, p_3), consistent estimators $\widehat{\mathrm{TCF}}_1(t_1; z)$, $\widehat{\mathrm{TCF}}_2(t_1, t_2; z)$, $\widehat{\mathrm{TCF}}_3(t_2; z)$ and $\widehat{\mathrm{ROCs}}(p_1, p_3; z)$ are straightforwardly obtained by the plug-in principle, i.e., substituting the REML

estimators $\widehat{\gamma}$ into expressions (3.1) and (3.2). The estimated covariate-specific ROC surface is visualized by plotting the points $(p_1, p_3, \widehat{\mathrm{ROCs}}(p_1, p_3; z))$ in the three-dimensional space.

It is straightforward to prove that the estimator $\left(\widehat{\mathrm{TCF}}_1(t_1; z), \widehat{\mathrm{TCF}}_2(t_1, t_2; z), \widehat{\mathrm{TCF}}_3(t_2; z)\right)^\top$ has an asymptotic normal distribution, by applying the delta method. Its asymptotic covariance matrix $\boldsymbol{\Omega}$ is

$$\boldsymbol{\Omega} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\gamma}^\top} \boldsymbol{\Lambda} \left(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\gamma}^\top}\right)^\top, \tag{3.3}$$

where $\boldsymbol{y} = (\mathrm{TCF}_1(t_1; z), \mathrm{TCF}_2(t_1, t_2; z), \mathrm{TCF}_3(t_2; z))$. Therefore, an approximate 95% confidence region for $\boldsymbol{y}$ has contour

$$(\boldsymbol{y} - \widehat{\boldsymbol{y}})^\top \widehat{\boldsymbol{\Omega}}^{-1} (\boldsymbol{y} - \widehat{\boldsymbol{y}}) = \chi^2_{0.95,3}, \tag{3.4}$$

where $\widehat{\boldsymbol{y}} = \left(\widehat{\mathrm{TCF}}_1(t_1; z), \widehat{\mathrm{TCF}}_2(t_1, t_2; z), \widehat{\mathrm{TCF}}_3(t_2; z)\right)$, $\widehat{\boldsymbol{\Omega}}$ is the estimate of $\boldsymbol{\Omega}$ and $\chi^2_{0.95,3}$ is the 95-th percentile of the χ^2 distribution with 3 degrees of freedom.

3.2 Selection of optimal thresholds

Under the considered model, a covariate-specific optimal pair of thresholds (t_1^+, t_2^+) for clustered data can be obtained by:

(i) maximizing the covariate-specific generalized Youden index $J_3(z)$ (GYI), with

$$J_3(z) = \mathrm{TCF}_1(t_1; z) + \mathrm{TCF}_2(t_1, t_2; z) + \mathrm{TCF}_3(t_2; z); \tag{3.5}$$

(ii) minimizing the covariate-specific Euclidean distance $D_3(z)$ between the ideal point $(1, 1, 1)$ and the point $(\mathrm{TCF}_1(t_1; z), \mathrm{TCF}_2(t_1, t_2; z), \mathrm{TCF}_3(t_2; z))$ (CtP), with

$$D_3(z) = \sqrt{[1 - \mathrm{TCF}_1(t_1; z)]^2 + [1 - \mathrm{TCF}_2(t_1, t_2; z)]^2 + [1 - \mathrm{TCF}_3(t_2; z)]^2}; \tag{3.6}$$

(iii) maximizing the covariate-specific volume $V_3(z)$ of the cuboid under the covariate-specific ROC surface (MV), with

$$V_3(z) = \mathrm{TCF}_1(t_1; z) \times \mathrm{TCF}_2(t_1, t_2; z) \times \mathrm{TCF}_3(t_2; z). \tag{3.7}$$

Observe that the covariate-specific objective functions $J_3(z)$, $D_3(z)$, $V_3(z)$ (and associated optimal pair of thresholds, say (t_1^+, t_2^+)) depend on $\boldsymbol{\gamma}$. Plugging the REML estimator $\widehat{\boldsymbol{\gamma}}$ into (3.5), (3.6) and (3.7), leads to the estimated versions $\widehat{J}_3(z)$, $\widehat{D}_3(z)$ and $\widehat{V}_3(z)$. Then, the estimators $(\hat{t}^+_{1,\mathrm{GYI}}, \hat{t}^+_{2,\mathrm{GYI}})$, $(\hat{t}^+_{1,\mathrm{MV}}, \hat{t}^+_{2,\mathrm{MV}})$ and $(\hat{t}^+_{1,\mathrm{CtP}}, \hat{t}^+_{2,\mathrm{CtP}})$ are obtained by maximizing $\widehat{J}_3(z)$ and $\widehat{V}_3(z)$, or minimizing $\widehat{D}_3(z)$,

under the constraint $t_1 < t_2$. The optimization leads also to the estimated covariate-specific optimal statistics $\widehat{J}_3^+(z)$, $\widehat{D}_3^+(z)$ and $\widehat{V}_3^+(z)$ (e.g., $\widehat{J}_3^+(z)$ is the maximum of $\widehat{J}_3(z)$).

The estimators $(\hat{t}^+_{1,\mathrm{GYI}}, \hat{t}^+_{2,\mathrm{GYI}})$, $(\hat{t}^+_{1,\mathrm{CtP}}, \hat{t}^+_{2,\mathrm{CtP}})$ and $(\hat{t}^+_{1,\mathrm{MV}}, \hat{t}^+_{2,\mathrm{MV}})$ are functions of the REML estimator $\widehat{\boldsymbol{\gamma}}$, but the function can be obtained in an explicit form for the GYI approach only:

$$\begin{aligned}
\hat{t}^+_{1,\mathrm{GYI}} &= \frac{\left(z^\top\widehat{\boldsymbol{\beta}}_2\widehat{\sigma}^2_{1c} - z^\top\widehat{\boldsymbol{\beta}}_1\widehat{\sigma}^2_{2c}\right) - \widehat{\sigma}_{1c}\widehat{\sigma}_{2c}\sqrt{\left(z^\top\widehat{\boldsymbol{\beta}}_1 - z^\top\widehat{\boldsymbol{\beta}}_2\right)^2 + (\widehat{\sigma}^2_{1c} - \widehat{\sigma}^2_{2c})\log\left(\frac{\widehat{\sigma}^2_{1c}}{\widehat{\sigma}^2_{2c}}\right)}}{\widehat{\sigma}^2_{1c} - \widehat{\sigma}^2_{2c}}, \\
\hat{t}^+_{2,\mathrm{GYI}} &= \frac{\left(z^\top\widehat{\boldsymbol{\beta}}_3\widehat{\sigma}^2_{2c} - z^\top\widehat{\boldsymbol{\beta}}_2\widehat{\sigma}^2_{3c}\right) - \widehat{\sigma}_{2c}\widehat{\sigma}_{3c}\sqrt{\left(z^\top\widehat{\boldsymbol{\beta}}_2 - z^\top\widehat{\boldsymbol{\beta}}_3\right)^2 + (\widehat{\sigma}^2_{2c} - \widehat{\sigma}^2_{3c})\log\left(\frac{\widehat{\sigma}^2_{2c}}{\widehat{\sigma}^2_{3c}}\right)}}{\widehat{\sigma}^2_{2c} - \widehat{\sigma}^2_{3c}},
\end{aligned}$$

where $\widehat{\sigma}^2_{ic} = \widehat{\sigma}^2_c + \widehat{\sigma}^2_i$, for $i = 1, 2, 3$ (see also [19]). However, in the Appendix we show that they are all consistent and asymptotically normal, with asymptotic covariance matrix

$$\boldsymbol{\Sigma}_{\hat{t}^+_{1,*},\hat{t}^+_{2,*}} = \left(\frac{\partial t^+_{1,*}}{\partial\boldsymbol{\gamma}^\top}, \frac{\partial t^+_{2,*}}{\partial\boldsymbol{\gamma}^\top}\right)\boldsymbol{\Lambda}\left(\frac{\partial t^+_{1,*}}{\partial\boldsymbol{\gamma}^\top}, \frac{\partial t^+_{2,*}}{\partial\boldsymbol{\gamma}^\top}\right)^\top, \tag{3.8}$$

where the symbol $*$ stands for the name of the selection method (i.e., GYI, CtP and MV) and

$$\frac{\partial t^+_{m,*}}{\partial\boldsymbol{\gamma}^\top} = \left(\frac{\partial^2 H}{\partial t^+_{m,*}\partial t^+_{m,*}}\right)^{-1}\left(-\frac{\partial^2 H}{\partial t^+_{m,*}\partial\boldsymbol{\gamma}^\top}\right),$$

$m = 1, 2$. The plug-in method gives consistent estimates of quantities in (3.8). Confidence regions can be easily constructed by using the normal approximation result.

3.3 Extension with Box–Cox transformation

In model (2.1), we assumed that α_k and ε_i follow a normal distribution. However, such assumption could be quite restrictive, and is violated in practical situations where data distribution may be skewed. In such situations, one solution is to resort to the Box-Cox transformation for linear mixed-effect models [14,20].

In particular, here we consider the model

$$\begin{aligned}
Y_1^{(\lambda)} &= \alpha_{k_1} + z_1^\top\boldsymbol{\beta}_1 + \varepsilon_1, \\
Y_2^{(\lambda)} &= \alpha_{k_2} + z_2^\top\boldsymbol{\beta}_2 + \varepsilon_2, \\
Y_3^{(\lambda)} &= \alpha_{k_3} + z_3^\top\boldsymbol{\beta}_3 + \varepsilon_3,
\end{aligned} \tag{3.9}$$

where $Y_i^{(\lambda)}$ is the transformed response

$$Y_i^{(\lambda)} = \begin{cases} \dfrac{Y_i^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \log\left(Y_i\right), & \lambda = 0 \end{cases}$$

with $i = 1, 2, 3$ ($Y_i > 0$), and λ is the transformation parameter[21]. Assumptions about the random effects α_k and the subject-level random errors ε_i are the same as in model (2.1). Therefore, the restricted log-likelihood function $\ell_R(\boldsymbol{\gamma}; \lambda)$ becomes

$$\begin{aligned} \ell_R(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta}; \lambda) \;=\; & -\frac{1}{2}\sum_{k=1}^{c}\left(\mathbf{Y}_k^{(\lambda)} - \mathbf{Z}_k\widehat{\boldsymbol{\beta}}_\lambda(\boldsymbol{\theta})\right)^{\top}\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{Y}_k^{(\lambda)} - \mathbf{Z}_k\widehat{\boldsymbol{\beta}}_\lambda(\boldsymbol{\theta})\right) - \frac{1}{2}\sum_{k=1}^{c}\log|\boldsymbol{\Sigma}_k| \\ & - \frac{1}{2}\log\left|\sum_{k=1}^{c}\mathbf{Z}_k^{\top}\boldsymbol{\Sigma}_k^{-1}\mathbf{Z}_k\right| + (\lambda - 1)\sum_{k=1}^{c}\sum_{i=1}^{3}\sum_{j=1}^{n_{ki}}\log(Y_{kij}), \end{aligned} \tag{3.10}$$

where $\mathbf{Y}_k^{(\lambda)}$ is the n_k-vector of the transformed responses within the cluster k, and

$$\widehat{\boldsymbol{\beta}}_\lambda(\boldsymbol{\theta}) \equiv \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}, \lambda) = \left(\sum_{k=1}^{c}\mathbf{Z}_k^{\top}\boldsymbol{\Sigma}_k^{-1}\mathbf{Z}_k\right)^{-1}\sum_{k=1}^{c}\mathbf{Z}_k^{\top}\boldsymbol{\Sigma}_k^{-1}\mathbf{Y}_k^{(\lambda)}.$$

Since direct maximization of (3.10) can produce unstable estimates of λ, we suggest to obtain $\widehat{\lambda}$ resorting to the method proposed by[22] and reviewed in Appendix. Once the estimate of λ has been obtained, $\ell_R(\boldsymbol{\gamma}; \widehat{\lambda})$ can be maximized to obtain $\widehat{\boldsymbol{\gamma}}$.

By using the estimates $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\lambda}$, it is straightforward to obtain covariate-specific estimates of points on the ROC surface, $\widehat{\mathrm{ROCs}}^{(\hat{\lambda})}(p_1, p_3; z)$, and to get estimated versions $\widehat{J}_3^{(\hat{\lambda})}(z)$, $\widehat{D}_3^{(\hat{\lambda})}(z)$ and $\widehat{V}_3^{(\hat{\lambda})}(z)$ of the covariate-specific objective functions $J_3^{(\lambda)}(z)$, $D_3^{(\lambda)}(z)$ and $V_3^{(\lambda)}(z)$, respectively. Then, the covariate-specific optimal pairs of thresholds for clustered data $\left(\hat{t}_{1,\mathrm{GYI}}^{+,(\hat{\lambda})}, \hat{t}_{2,\mathrm{GYI}}^{+,(\hat{\lambda})}\right)$, $\left(\hat{t}_{1,\mathrm{CtP}}^{+,(\hat{\lambda})}, \hat{t}_{2,\mathrm{CtP}}^{+,(\hat{\lambda})}\right)$ and $\left(\hat{t}_{1,\mathrm{MV}}^{+,(\hat{\lambda})}, \hat{t}_{2,\mathrm{MV}}^{+,(\hat{\lambda})}\right)$ are derived by maximizing the corresponding objective functions in the transformation scale. Therefore, by the inversion

$$t = \begin{cases} \left(\lambda t^{(\lambda)} + 1\right)^{1/\lambda}, & \lambda \neq 0 \\ \exp\left(t^{(\lambda)}\right), & \lambda = 0, \end{cases} \tag{3.11}$$

the covariate-specific optimal pair of thresholds in the original scale can be obtained ($t^{(\lambda)}$ denotes a value in the transformed scale).

All the above estimators are consistent and asymptotically normally distributed. The asymptotic covariance matrix can be obtained and estimated by employing the delta method and the plug-in principle. However, in our experience, such an approach can lead to unstable estimates. For this reason, we advise to use in such context a nonparametric bootstrap procedure for clustered data, which is described in the following steps (taking as an example of parameter of interest the pair of optimal thresholds):

- Step 1. Obtain the estimates $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\lambda}$ in the transformed data scale. Then, estimate the covariate-specific optimal pair of thresholds $\left(\hat{t}^{+,(\hat{\lambda})}_{1,*}, \hat{t}^{+,(\hat{\lambda})}_{2,*}\right)$, and use (3.11) to go back to $\left(\hat{t}^{+}_{1,*}, \hat{t}^{+}_{2,*}\right)$.
- Step 2. Draw c clusters, with replacement, from the set of clusters; then pick up all observations within the sampled clusters.
- Step 3. Based on the sample generated in Step 2, obtain the estimated parameters $\widehat{\boldsymbol{\gamma}}_{(b)}$ and $\widehat{\lambda}_{(b)}$. Then, estimate the covariate-specific optimal pair of thresholds $\left(\hat{t}^{+,(\hat{\lambda})}_{1,*(b)}, \hat{t}^{+,(\hat{\lambda})}_{2,*(b)}\right)$ and transform back to the original scale to get $\left(\hat{t}^{+}_{1,*(b)}, \hat{t}^{+}_{2,*(b)}\right)$.
- Step 4: Repeat steps 2 and 3 B times, and compute the bootstrap-based estimate of covariance matrix of $\left(\hat{t}^{+}_{1,*}, \hat{t}^{+}_{2,*}\right)$ as

$$\frac{1}{B-1}\sum_{b=1}^{B}\left\{\begin{pmatrix}\hat{t}^{+}_{1,*(b)}\\ \hat{t}^{+}_{2,*(b)}\end{pmatrix}-\begin{pmatrix}\bar{\hat{t}}^{+}_{1,*}\\ \bar{\hat{t}}^{+}_{2,*}\end{pmatrix}\right\}\left\{\begin{pmatrix}\hat{t}^{+}_{1,*(b)}\\ \hat{t}^{+}_{2,*(b)}\end{pmatrix}-\begin{pmatrix}\bar{\hat{t}}^{+}_{1,*}\\ \bar{\hat{t}}^{+}_{2,*}\end{pmatrix}\right\}^{\top},$$

where $\bar{\hat{t}}^{+}_{1,*}=\frac{1}{B}\sum_{b=1}^{B}\hat{t}^{+}_{1,*(b)}$ and $\bar{\hat{t}}^{+}_{2,*}=\frac{1}{B}\sum_{b=1}^{B}\hat{t}^{+}_{2,*(b)}$.

Recall that symbol $*$ stands for the name of the selection criterion (i.e., GYI, CtP and MV).

It is important to emphasize that the approach discussed in this subsection can only solve problems arising from particular forms of violation of the assumption of normality. Indeed, the motivating idea is that the same transformation is suitable for all classes of disease. In case of more complex deviations from normality, it is necessary to resort to some completely non-parametric method. This topic is outside the scope of this paper and deserves future work.

4 Simulation study

4.1 Simulation set-up

We perform several simulation experiments to evaluate the performance of the proposed estimators of the covariate-specific ROC surface and the optimal pair of thresholds. In all simulations, the number of clusters c is taken to belong to the set $\{15, 30, 60\}$, and in k-th cluster, the disease status for subjects is generated from a multinomial distribution, $Mult(n_k, (0.6, 0.3, 0.1))$. We consider three different settings, two in the Gaussian setting and one that requires the Box-Cox transformation.

- Setting 1. We consider one covariate $X \sim \mathcal{U}(-2, 2)$. The parameters of model (2.1) are set to be $\beta_{01} = 0.5$, $\beta_{11} = 0.5$, $\beta_{02} = 2$, $\beta_{12} = 0.8$, $\beta_{03} = 3.5$ and $\beta_{13} = 1.1$. Variances of errors ε_i are set to be $\sigma_1^2 = 0.3$, $\sigma_2^2 = 0.8$, $\sigma_3^2 = 1.3$; σ_c^2 is set to be 0.2 or 1. Some true covariate-specific VUS values are roughly 0.505 at $x = -2$, 0.733 at $x = 0.11$ and 0.867 at $x = 2$, when $\sigma_c^2 = 0.2$; 0.415 at $x = -2$, 0.605 at $x = 0.11$ and 0.746 at $x = 2$, when $\sigma_c^2 = 1$. Just as an example, Figure S16 in Supplementary Material shows the corresponding true covariate-specific ROC surfaces, when $\sigma_c^2 = 0.2$.
- Setting 2. We consider two covariates, $X_1 \sim \mathcal{N}(0, 1)$ and $X_2 \sim \mathcal{B}\text{ernoulli}(0.5)$. The parameters of model (2.1) are set to be $\beta_{01} = -0.5$, $\beta_{11} = 0.5$, $\beta_{21} = -0.5$, $\beta_{02} = 2$, $\beta_{12} = 1$, $\beta_{22} = -0.2$, $\beta_{03} = 3$, $\beta_{13} = 1.5$ and $\beta_{23} = 0.6$. Variances of errors ε_i are set to be $\sigma_1^2 = 0.5$, $\sigma_2^2 = 1$, $\sigma_3^2 = 1.5$; σ_c^2 is set to be 0.3 or 1.4. Some true covariate-specific VUS values are roughly 0.379 at $(x_1, x_2) = (-2, 0)$, 0.689 at $(x_1, x_2) = (0, 11, 0)$, 0.866 at $(x_1, x_2) = (2, 0)$, 0.578 at $(x_1, x_2) = (-2, 1)$, 0.831 at $(x_1, x_2) = (0, 11, 1)$, 0.941 at $(x_1, x_2) = (2, 1)$, when $\sigma_c^2 = 0.3$; 0.325 at $(x_1, x_2) = (-2, 0)$, 0.584 at $(x_1, x_2) = (0, 11, 0)$, 0.771 at $(x_1, x_2) = (2, 0)$, 0.476 at $(x_1, x_2) = (-2, 1)$, 0.717 at $(x_1, x_2) = (0, 11, 1)$, 0.861 at $(x_1, x_2) = (2, 1)$, when $\sigma_c^2 = 1.4$.
- Setting 3. We consider one covariate $X \sim \mathcal{U}(0.5, 2)$. The parameters of model (3.9) are set to be $\beta_{01} = 2$, $\beta_{11} = 2$, $\beta_{02} = 3$, $\beta_{12} = 3.5$, $\beta_{03} = 3.5$ and $\beta_{13} = 3$. Variances of errors ε_i are set to be $\sigma_1^2 = 0.3$, $\sigma_2^2 = 0.48$, $\sigma_3^2 = 0.84$; σ_c^2 is set to be 0.16. We fix the transformation parameter λ as 0.5. Some true covariate-specific VUS values are roughly 0.636 at $x = 0.5$, 0.758 at $x = 1.16$ and 0.864 at $x = 2$.

The choice of variance components allow us to fix different values of the intra-class correlation coefficient (ICC), defined as $ICC = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_\varepsilon^2}$, with $\sigma_\varepsilon = \frac{1}{3}(\sigma_1 + \sigma_2 + \sigma_3)$. In particular, the ICC equals 0.213 or 0.574 in our Setting 1, 0.239 or 0.594 in Setting 2, and 0.236 in Setting 3. The diagnostic test results are generated according to model (2.1) in Settings 1 and 2, and model (3.9) in Setting 3. We consider two cases for the sample size within cluster: (i) balanced design, i.e., all clusters have the same size, with $n_k \in \{4, 10\}$; (ii) unbalanced design, i.e., each cluster has different size n_k, (randomly) varying from 3 to 14 (the latter case is the only one considered in Setting 3). For each simulation experiment, the number of Monte Carlo replications is 1000.

Clearly, for each considered scenario, at each Monte Carlo replication, the covariate-specific ROC surface and the covariate-specific optimal pair of thresholds are estimated according to the methods proposed in Section 3. The estimated variances of the covariate-specific estimators for the optimal thresholds are obtained by applying the plug-in method to (3.8) in Settings 1 and 2, and by using the cluster bootstrap procedure, with 200 bootstrap replication, in Setting 3. According to our monotone ordering assumption, only consistent generated samples (or bootstrap samples), i.e. such that $z^\top \widehat{\boldsymbol{\beta}}_1 <$

$z^\top \widehat{\boldsymbol{\beta}}_2 < z^\top \widehat{\boldsymbol{\beta}}_3$, are processed. This selection is required by the statistical tools considered in the paper, that are valid, from a methodological point of view, under the monotone ordering assumption.[7,9]

4.2 Results for the covariate-specific ROC surface

The covariate-specific ROC surface estimator is evaluated, in every setting, by means of the Monte Carlo integrated mean bias (Bias) and the Monte Carlo square root of the integrated mean squared error (RMSE), i.e.,

$$\text{Bias}\left(\widehat{\text{ROCs}}(p_1,p_3;z)\right) = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{n_{p_1}n_{p_3}}\sum_{i=1}^{n_{p_1}}\sum_{j=1}^{n_{p_3}}\left(\widehat{\text{ROCs}}(p_{1i},p_{3j};z) - \text{ROCs}(p_{1i},p_{3j};z)\right)$$

and

$$\text{RMSE}\left(\widehat{\text{ROCs}}(p_1,p_3;z)\right) = \left\{\frac{1}{M}\sum_{m=1}^{M}\frac{1}{n_{p_1}n_{p_3}}\sum_{i=1}^{n_{p_1}}\sum_{j=1}^{n_{p_3}}\left(\widehat{\text{ROCs}}(p_{1i},p_{3j};z) - \text{ROCs}(p_{1i},p_{3j};z)\right)^2\right\}^{1/2},$$

where $M = 1000$ is the total number of Monte Carlo replications, n_{p_1} and n_{p_3} are the number of grid points we used for $p_1(= TCF_1)$ and $p_3(= TCF_3)$, respectively. More precisely, we set $n_{p_1} = n_{p_3} = 21$.

Results for Setting 1 are shown in Figure 1 (Figures S1, and S9 in Supplementary Material give simulation results for Settings 2 and 3, respectively). We can see that larger values of Bias and RMSE are present in cases of small number of clusters or sample size within clusters. As expected, increasing sample sizes improves the accuracy of the covariate-specific ROC surface's estimator.

We also consider the problem of constructing (joint) confidence regions for the covariate-specific true class fractions at a fixed pair of thresholds, for instance $t_1 = 0.5$ and $t_2 = 3.5$, in Setting 1. In this case, Figure 2 (and S8, S13 in Supplementary Material) shows an evident liberal behavior, with low coverage in case of smallest within-cluster sample size ($n_k = 4$) and/or smallest number of clusters ($c = 15$). However, the empirical coverages increase when either the number of clusters or the within-cluster sizes increase. For comparison purposes, we also performed the Naïve estimators for the covariate-specific true class fractions, which assume independence of all subjects and ignore the cluster-level effect. Results in Figure 2 (and S8, S13 in Supplementary Material) show the effect of this misconception when building confidence regions using pivots based on those estimators: the actual coverage level stays away from the nominal one, even when the sample size becomes large.

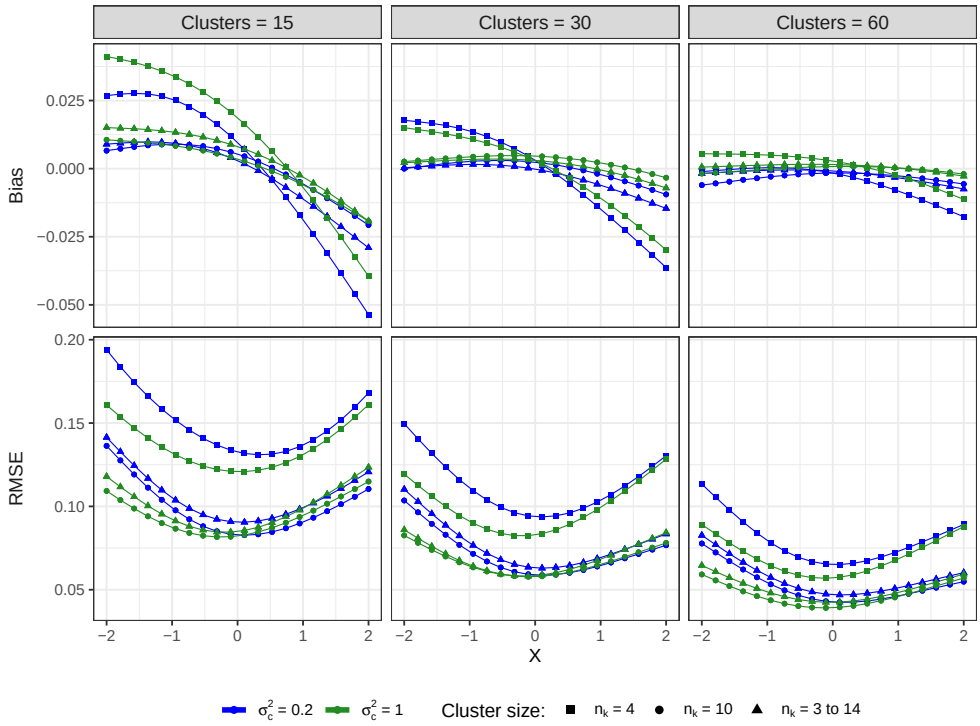

Figure 1. Monte Carlo Bias and RMSE of the estimator for the covariate-specific ROC surface in Setting 1.

4.3 Results for the optimal pair of thresholds

The covariate-specific proposed estimators are evaluated in terms of Bias (difference between Monte Carlo mean and the truth), root mean square error (RMSE: the square root of the sum of squared bias and Monte Carlo variance), and coverage probability (CP) of 95% confidence intervals (obtained by using the normal approximation approach).

Figures 3 and 4 present Bias and RMSE, respectively, of the covariate-specific optimal pair of thresholds estimators, as a function of covariate values across all simulated scenarios for Setting 1. Figure 5 shows empirical coverages of corresponding 95% confidence regions for the true parameters. For space reasons, we report the results for Settings 2 and 3 in Supplementary Material.

As shown in Figure 3 (also in Figures S2, S3 and S10 in Supplementary Material), Bias depends on the total number of clusters c and on the within-cluster sample size n_k. When increasing either n_k or c, the bias decreases, and at the largest value of total number of clusters $c = 60$, the bias is close to zero

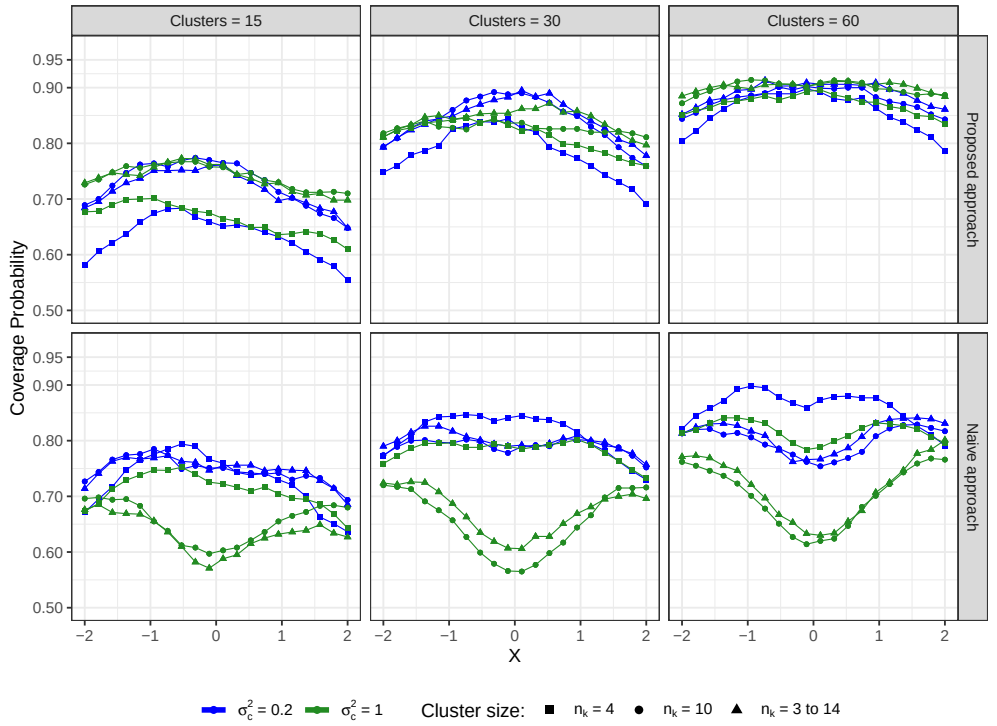

Figure 2. Empirical coverage of the (joint) confidence regions for the true class fractions, at $t_1 = 0.5$ and $t_2 = 3.5$, build by using the proposed three selection methods and the Naïve approach (with normal approximation), in Setting 1. Nominal level 0.95.

regardless of the within-cluster sample size n_k, although a slightly large bias can be observed in cases of small cluster sizes, when estimating Threshold 2. The ICC also seems to affect the bias; in particular, larger values of ICC lead to larger values of Bias.

In terms of RMSE, we can see similar trends: RMSE decrease when total number of clusters and/or within-cluster size increases. Figure 4 (and S4, S5, S11 in Supplementary Material) also clearly shows the impact of ICC on the RMSE, especially in cases of small cluster size. It implies that larger cluster variance σ_c^2 is associated with larger variances of the covariate-specific optimal pair of thresholds estimators.

In general, we notice that the three discussed selection methods behave similarly, and that large values of Bias and RMSE of the threshold estimators appear at some ranges of covariate values (typically in the

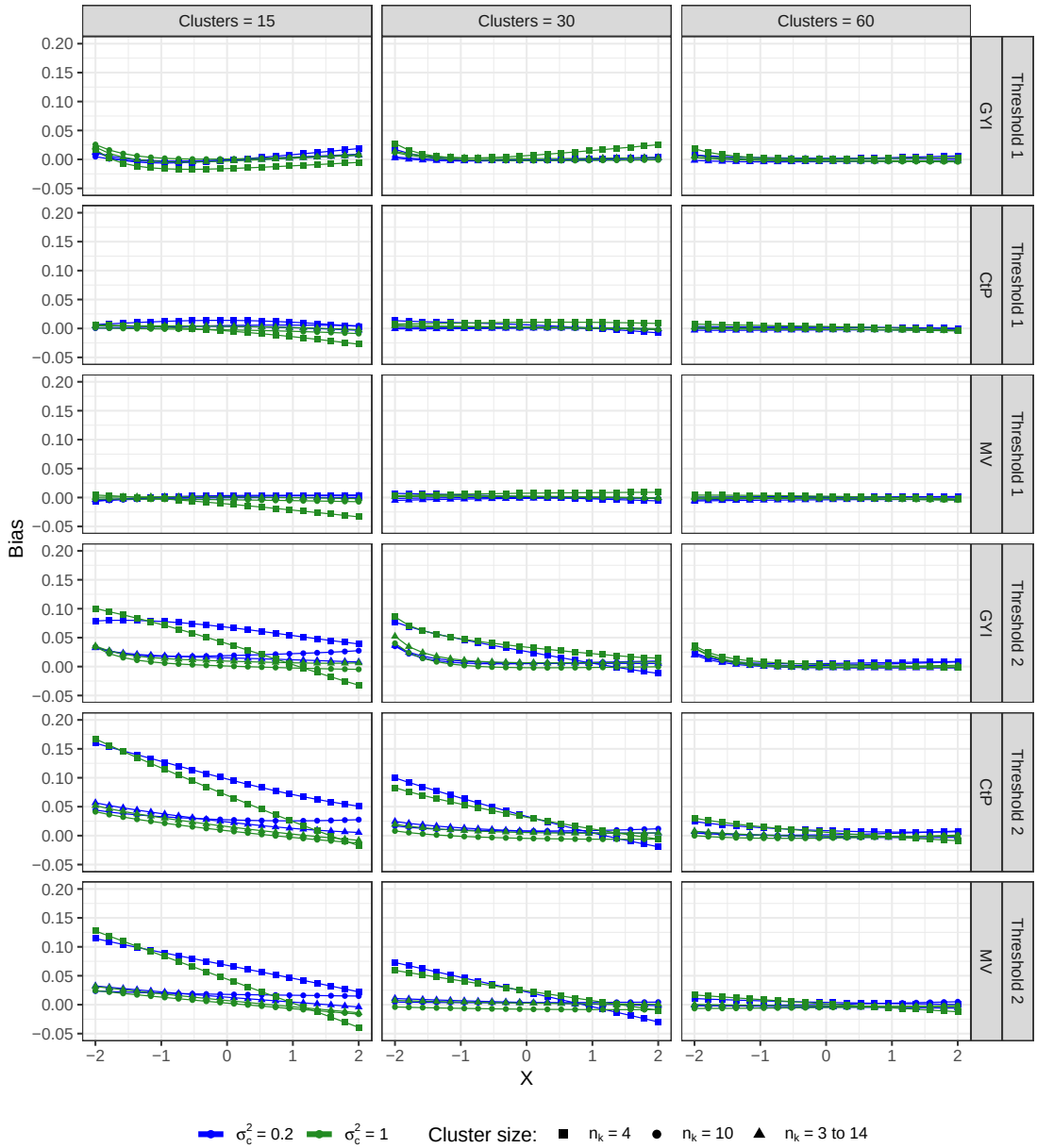

Figure 3. Monte Carlo Bias of covariate-specific optimal pair of thresholds estimators, obtained by the three proposed selection methods in Setting 1.

border area), in case of small sample sizes. This may be due to a large degree of overlap of the covariate-specific test distributions (in two contiguous classes) around the true optimal thresholds, in combination with a small number of observations, that not allow accurate estimation.

We also consider the problem of constructing (joint) confidence regions for the covariate-specific optimal thresholds. In this case, Figure 5 (and S6, S7, S12 in Supplementary Material) again shows a certain liberality of the regions, having these low coverage in case of smallest within-cluster sample size ($n_k = 4$) and smallest number of clusters ($c = 15$). However, the empirical coverages increase when the sample size (globally) increases. For comparison purposes, we also employed the Naïve estimators for the covariate-specific optimal pair of thresholds. Results in Figure 5 (and S6, S7, S12 in Supplementary Material) show the dramatic effect of the misconception when building confidence regions using pivots based on those estimators.

5 Application

Here, we show how to apply the proposed method to the problem of celltype classification, introduced in Section 1. Although RNA sequencing provides a comprehensive view of the cells, measuring thousands of genes at the time, the goal is often to identify a handful of biomarkers to use in low-throughput experiments (such as in situ hybridization) that allow to visualize the presence of the gene through fluorescence microscopes. In this context, it is useful to test in RNA sequencing data the predictive role of genes to discriminate cell types. Here, we focus on the *Lysosomal Associated Membrane Protein Family Member 5* (Lamp5) gene and on its ability to discriminate three types of glutamatergic neurons, namely Layer 2/3 Intratelencephalic (L2/3 IT), Layer 4 (L4) and Layer 5 Pyramidal Tract (L5 PT) neurons.

Overall, our dataset consists of 860 observations (cells): 265 in the L2/3 IT group, 152 in the L4 group, and 443 in the L5 PT group. For each observation, the following variables were measured: the expression of the Lamp5 gene (biomarker), the mouse genotype (which yields 23 clusters), the class labels (L2/3 IT, L4, and L5 PT), and the sex and age (in days) of the mouse. The sample size within cluster varies from 1 to 330, with 20 clusters ranging from 1 to 50 cells, and 3 clusters consisting more than 100 cells. As, in this case, the rank-ordered nature of the biomarker with respect to the classes is not given, the monotone ordering was specified by ordering the classes according to the rank of the biomarker's sample means in the three groups.

The expression of the biomarker is measured through RNA sequencing, which yields count data, rescaled to count per million (CPM) to account for the differences in sequencing depth between samples [15]. In order to make the normality assumption likely, we thus consider the linear mixed-effect model under the Box-Cox transformation (3.9) for the CPM of Lamp5 using as the covariate the age (in days) of the mice. The estimated Box-Cox parameter is 0.446. The estimated regression coefficients are

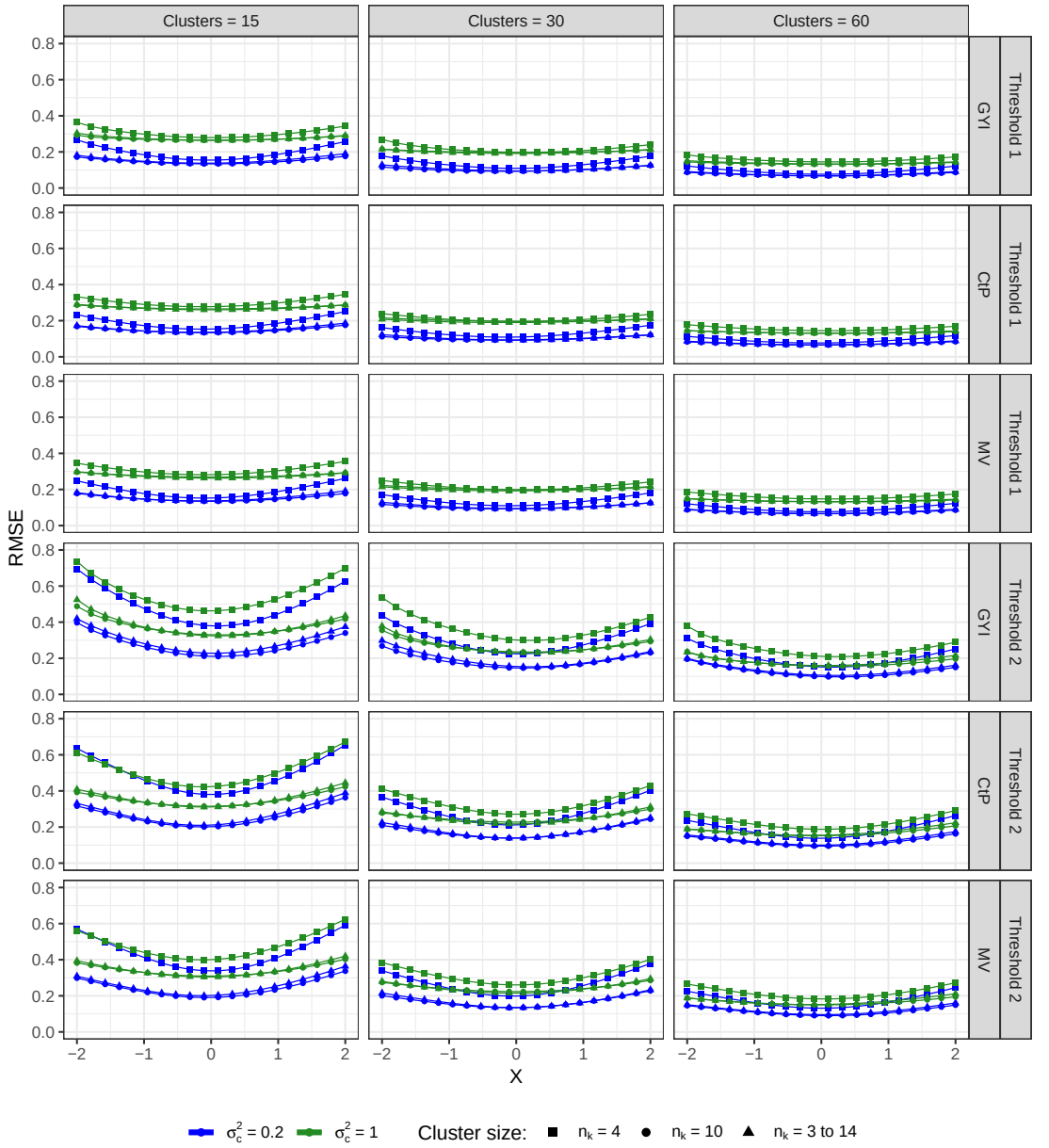

Figure 4. Monte Carlo RMSE of covariate-specific optimal pair of thresholds estimators obtained by the three proposed selection methods in Setting 1.

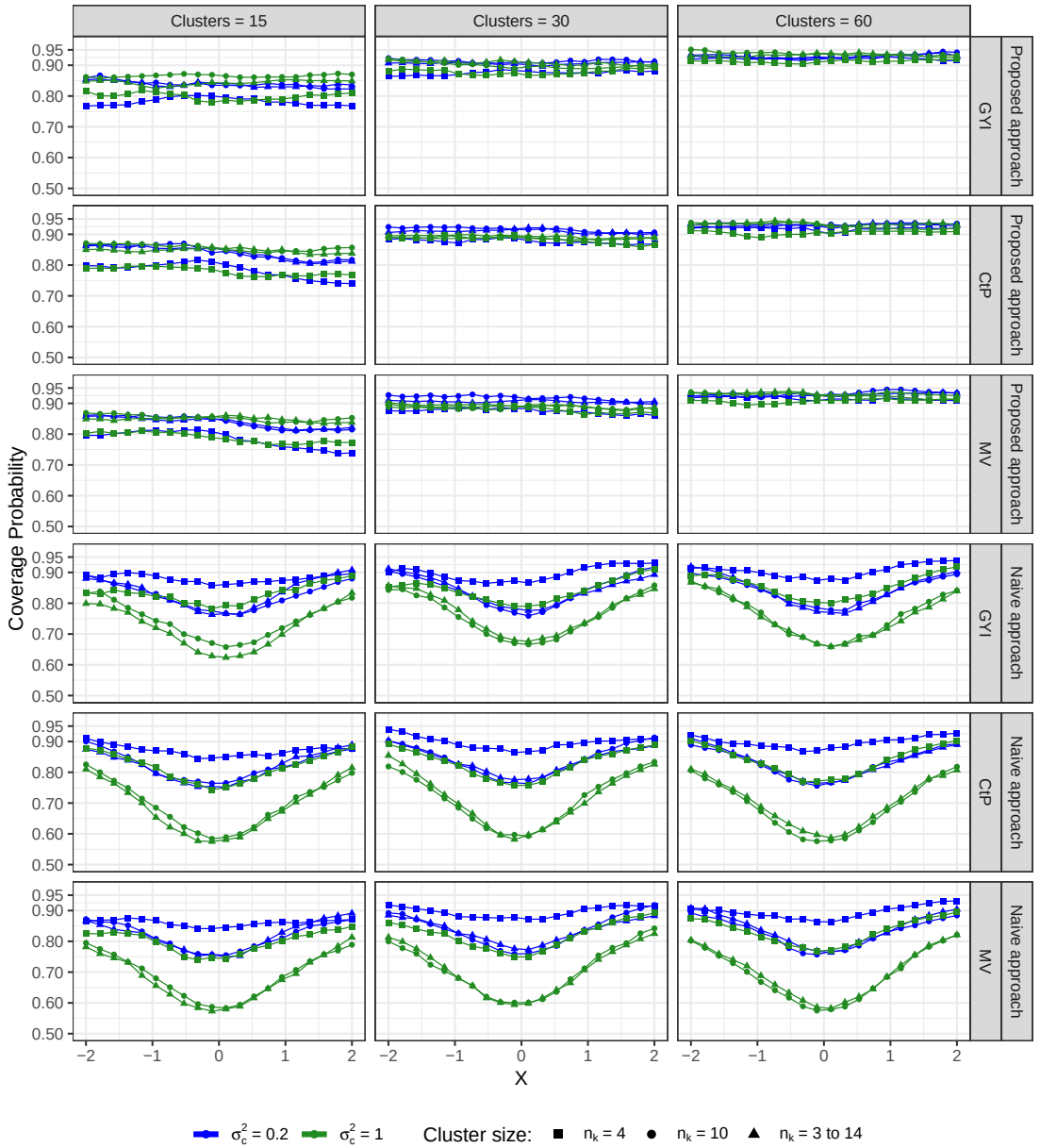

Figure 5. Empirical coverage of the (joint) confidence regions for the true optimal thresholds, build by using the proposed three selection methods and the Naïve approach (with normal approximation), in Setting 1. Nominal level 0.95.

reported in Table 1. Significant positive relationship between the transformed response (CPM of Lamp5) and the age was found for class L4 and class L5 PT. The estimated ICC is about 0.229, which indicates a weak correlation between observations in a same cluster (Genotype ID). The verification of the model assumptions is given by Figures S14 and S15, Supplementary Material.

Table 1. Estimated parameters in the linear mixed model for CPM of Lamp5 under the Box-Cox transformation.

	Covariates	Estimate	Standard error	p-value
Class L4	Intercept	0.788	5.953	0.895
	Age	0.450	0.182	0.013
Class L5 PT	Intercept	34.305	12.280	0.005
	Age	0.210	0.107	0.049
Class L2/3 IT	Intercept	49.346	20.882	0.018
	Age	0.090	0.083	0.279
Standard deviations	σ_c	6.786	–	–
	σ_1	15.025	–	–
	σ_2	11.241	–	–
	σ_3	11.143	–	–

Using the VUS estimator[12], we then proceeded to compute VUS values for specific age values based on the fitted model. Figure 6 displays the estimated VUS and corresponding 95% point-wise confidence band with the age (in day) of mouse varying from 51 to 84 days. The estimated VUS ranging between 0.40 to 0.56 indicates a quite good diagnostic ability of Lamp5. However, the accuracy decreases when the mouse gets older.

After estimating the age-specific VUS, we then apply our proposed methods to obtain the age-specific ROC surface at 54 days (Figure 7) and to obtain the optimal pair of thresholds based on the fitted model. Figure 8 presents the age-specific optimal pair of thresholds (t_1^+, t_2^+) and the corresponding 95% confidence regions, obtained by using three selection methods, GYI, CtP and MV. In order to obtain the confidence regions, we applied the cluster bootstrap procedure as mentioned in Section 3.2, with 200 bootstrap replications.

As shown in Figure 8, the age-specific optimal pair of thresholds increase with age linearly. Two selection methods, CtP and MV, yielded very similar values of age-specific optimal pair of thresholds estimates and confidence regions. In contrast, the GYI method produced smaller values for the estimates, and larger confidence regions.

We also tried to fit the linear mixed-effect model under the Box-Cox transformation (3.9) for the CPM of Lamp5 where the covariates include the age (in days) and the sex of the mice. However, we could not find any significant relationship between the transformed response and sex.

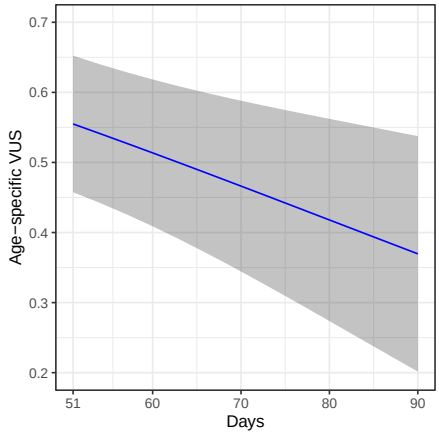

Figure 6. Age-specific estimates of VUS and corresponding 95% point-wise confidence intervals.

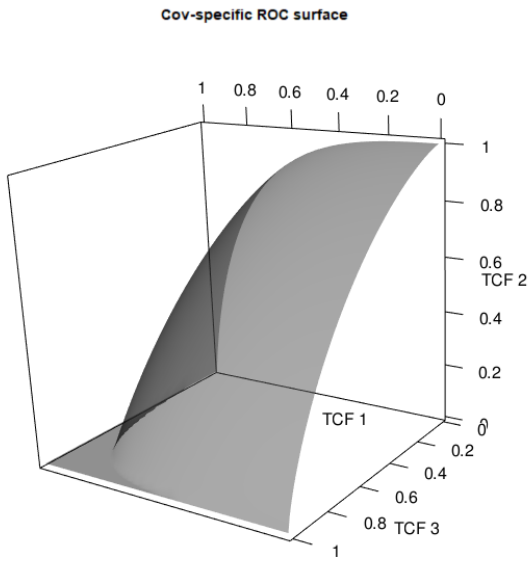

Figure 7. Age-specific estimated ROC surface at 54 days.

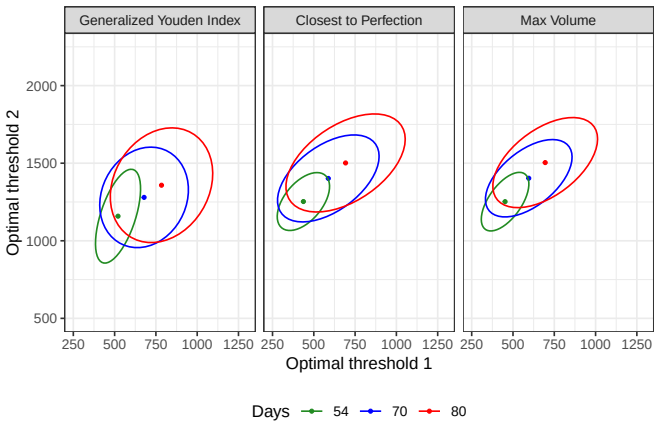

Figure 8. 95% confidence regions for the age-specific optimal thresholds for three values of age.

6 Discussion

In a three-class setting, the ROC surface represents a natural generalization of the ROC curve and is commonly used to evaluate the ability of a diagnostic test to distinguish among three classes (or levels) of a disease or, more generally, among three conditions to be classified.

In this paper, we have studied the problem of estimating covariate-specific ROC surfaces with clustered data, and derived three alternative estimators for the optimal threshold values for the test. For these new estimators, which find their rationale in three different approaches (GYI, CtP and MV), we have shown consistency and asymptotic normality, both under the assumption of normality, and in a more general context in which this assumption is (not heavily) violated. Asymptotic results can be used to build adequate covariate-specific confidence regions.

Our simulation results agree with the theoretical ones, and show a substantial behavioural equivalence for the three covariate-specific optimal thresholds estimators. In particular, confidence regions built using pivots based on the proposed estimators show a coverage level close to the nominal one, at least in medium to high sample sizes. From a practical point of view, the GYI approach has the advantage of providing estimates $(t^+_{1,\mathrm{GYI}}, t^+_{2,\mathrm{GYI}})$ in explicit form. However, it requires some attention as it may suffer from greater variability, particularly in certain regions (boundary regions) of the covariate space. From a computational point of view, the estimates $(t^+_{1,\mathrm{GYI}}, t^+_{2,\mathrm{GYI}})$ can be used as starting points for the procedures of numerical optimization that produce estimates according to the other two approaches (CtP, MV). Generally speaking, the large bias or poor coverage that we observed in small samples are essentially determined, in our opinion, by a lack of information contained in the samples. Just as

an example, figures S20-S22, in Supplementary Material, show the improvement in terms of bias (for covariate-specific ROC surface) and empirical coverage probability (of the joint confidence regions for the true class fractions) in Setting 1 of our simulations, when we consider balanced classes, i.e. the disease status is generated from a multinomial distribution, $Mult(n_k, (1/3, 1/3, 1/3))$.

When approaching statistical evaluation of a diagnostic test or biomarker, one typically has an idea about the possible association between the test and the disease status, so that elicitation of a monotone ordering for the classes may not represent a major criticality. In case this knowledge is too vague, as suggested by[9], pairwise AUCs for adjacent classes might also be used to reveal the correct order, which in turn can be used for the three-class analysis. However, in our paper, as well as in other papers dealing with covariate-specific ROC analysis, methods are designed under the assumption that the monotone ordering hypothesis holds for every value of the covariates. This assumption may not be satisfied in practice and its check might add, from a practical point of view, some complexity to the analysis. A rather pragmatic solution is based on the selection of the values for the covariates that are considered most interesting in terms of the diagnostic task, followed by a check on the corresponding covariate-specific VUS[12] estimated value, which is expected to be greater than 1/6. This exploratory analysis should be seen as a starting point for further investigation into the reasons for unusual results or apparent reversal of ordering.

Our proposal found a favourable response in the application. We have shown that a single gene, Lamp5, is able to well discriminate L2/3 IT, L4, and L5 PT glutamatergic neurons. However, more than one gene may be needed for more complex problems. For instance, Lamp5 is also expressed in a subset of GABAergic neurons; the joint distribution of Lamp5 and a glutamatergic biomarker (such as the Slc17a7 gene) may allow us to discriminate between Lamp5 GABAergic neurons and L2/3 IT, L4, and L5 PT neurons. Considering that RNA sequencing yields expression data for thousands of genes at a time, a multivariate, and possibly high-dimensional, version of the proposed model is of interest for molecular biology classification and will be considered in future research.

To conclude, we respond to the solicitation of a Reviewer who mentions a well-known intrinsic problem of linear mixed models and REML estimators, related to estimation of the variance of the random components. Indeed, estimation may suffer from poor accuracy, with estimates that can also result to be very close to zero. In our experience, also corroborated by simulation experiments not shown in the paper, very small values of the estimate of σ_c^2 (or other components of variance) seem to have a very limited impact on the inferential procedures proposed in our work. This evidence is illustrated in some graphs (figures S17-S19) presented in Supplementary Material.

Acknowledgements

The authors acknowledge the Associated Editor and two anonymous Reviewers whose valuable suggestions contributed to improve presentation of the contents.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Ministero dell'Istruzione, dell'Università e della Ricerca-Italy (grant number DIFO_ECCELLENZA18_01).

Supplemental material

Supplemental material for this article is available online.

References

1. Pepe MS. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, 2003.
2. Zhou XH, McClish DK and Obuchowski NA. *Statistical methods in diagnostic medicine*. John Wiley & Sons, 2011.
3. To Duc K, Chiogna M and Adimari G. Bias-corrected methods for estimating the receiver operating characteristic surface of continuous diagnostic tests. *Electron J Stat* 2016; 10(2): 3063–3113.
4. Lo K, Brinkman RR and Gottardo R. Automated gating of flow cytometry data via robust model-based clustering. *Cytometry Part A* 2008; 73(4): 321–332.
5. Ecker JR, Geschwind DH, Kriegstein AR et al. The brain initiative cell census consortium: lessons learned toward generating a comprehensive brain cell atlas. *Neuron* 2017; 96(3): 542–557.
6. Nakas CT and Yiannoutsos CT. Ordered multiple-class ROC analysis with continuous measurements. *Stat Med* 2004; 23(22): 3437–3449.
7. Nakas CT and Alonzo TA. ROC graphs for assessing the ability of a diagnostic marker to detect three disease classes with an umbrella ordering. *Biometrics* 2007; 63(2): 603–609.
8. 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. *Stat Med* 2010; 29(28): 2946–2955.

9. Nakas CT. Developments in ROC surface analysis and assessment of diagnostic markers in three-class classification problems. *REVSTAT – Stat J* 2014; 12(1): 43–65.
10. Attwood K, Tian L and Xiong C. Diagnostic thresholds with three ordinal groups. *J Biopharm Stat* 2014; 24(3): 608–633.
11. Hua J and Tian L. A comprehensive and comparative review of optimal cut-points selection methods for diseases with multiple ordinal stages. *J Biopharm Stat* 2020; 30(1): 46–68.
12. Xiong C, Luo J, Chen L et al. Estimating diagnostic accuracy for clustered ordinal diagnostic groups in the three-class case–Application to the early diagnosis of Alzheimer disease. *Stat Methods Med Res* 2018; 27(3): 701–714.
13. McCulloch CE and Searle SR. *Generalized, Linear, and Mixed Models*. John Wiley & Sons, 2001.
14. Gurka MJ, Edwards LJ, Muller KE et al. Extending the Box–Cox transformation to the linear mixed model. *J R Stat Soc A Stat* 2006; 169(2): 273–288.
15. Tasic B, Yao Z, Graybuck LT et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature* 2018; 563(7729): 72–78.
16. Liang KY and Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; 73(1): 13–22.
17. Kauermann G and Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. *J Am Stat Assoc* 2001; 96(456): 1387–1396.
18. Mancl LA and DeRouen TA. A covariance estimator for GEE with improved small-sample properties. *Biometrics* 2001; 57(1): 126–134.
19. Luo J and Xiong C. Youden index and associated cut-points for three ordinal diagnostic groups. *Commun Stat-Theor Simul C* 2013; 42(6): 1213–1234.
20. Lipsitz SR, Ibrahim J and Molenberghs G. Using a Box–Cox transformation in the analysis of longitudinal data with incomplete responses. *J R Stat SOC C-Appl* 2000; 49(3): 287–296.
21. Box GE and Cox DR. An analysis of transformations (with discussion). *J R Stat SOC B* 1964; 26(2): 211–252.
22. Gurka MJ and Edwards LJ. Estimating variance components and random effects using the box-cox transformation in the linear mixed model. *Commun Stat-Theor M* 2011; 40(3): 515–531.

Appendix

Consistency and asymptotic normality of $(\hat{t}^{+}_{1,*}, \hat{t}^{+}_{2,*})$

For a given vector of covariate values z, let $H(t_1, t_2, \boldsymbol{\gamma}; z)$ be the objective functions, i.e., $J_3(z)$, $D_3(z)$ or $V_3(z)$. Let $f(t_1, t_2, \boldsymbol{\gamma}; z) = \frac{\partial H}{\partial (t_1, t_2)^{\top}}$. The true and unknown optimal pair of thresholds, $(t^{+}_{1,0}, t^{+}_{2,0})$ say, maximizes or minimizes the function $H(t_1, t_2, \boldsymbol{\gamma}_0; z)$, or equivalently solves the equation

$f(t_1, t_2, \boldsymbol{\gamma}_0; z) = \mathbf{0}$, when $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$ is the true parameter value. Then, we have $f(t^+_{1,0}, t^+_{2,0}, \boldsymbol{\gamma}_0; z) = \mathbf{0}$ and

$$\left.\frac{\partial f(t_1, t_2, \boldsymbol{\gamma}_0; z)}{\partial (t_1, t_2)^\top}\right|_{(t_1,t_2)=(t^+_{1,0}, t^+_{2,0})}$$

has non-zero determinant, or equivalently, is invertible. Note that, the function f is continuously differentiable. By the implicit function theorem, there exists a neighborhood of $\boldsymbol{\gamma}_0$ where a unique continuously differentiable function $m(\boldsymbol{\gamma})$ is defined, such that $m(\boldsymbol{\gamma}_0) = (t^+_{1,0}, t^+_{2,0})$ and $f(m(\boldsymbol{\gamma}), \boldsymbol{\gamma}; z) = 0$ for all $\boldsymbol{\gamma}$ in the neighborhood of $\boldsymbol{\gamma}_0$. Since the REML estimator $\widehat{\boldsymbol{\gamma}}$ is consistent, i.e., $\widehat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}_0$, we have that $(\hat{t}^+_1, \hat{t}^+_2) = m(\widehat{\boldsymbol{\gamma}}) \xrightarrow{p} (t^+_{1,0}, t^+_{2,0})$, by using the continuous mapping theorem. Applying the Delta method, asymptotic normality of $(\hat{t}^+_1, \hat{t}^+_2)$ follows, with asymptotic covariance matrix given in (3.8).

Observe that also the covariate-specific estimator of the optimal value for the criterion function, $\widehat{H}^+(z) = H(\hat{t}^+_1, \hat{t}^+_2, \widehat{\boldsymbol{\gamma}}; z)$, is consistent and asymptotically normal, with covariance matrix

$$\mathbb{V}\mathrm{ar}\left(\widehat{H}^+\right) \quad = \quad \left(\frac{\partial H}{\partial \boldsymbol{\gamma}^\top}\right) \boldsymbol{\Lambda} \left(\frac{\partial H}{\partial \boldsymbol{\gamma}^\top}\right)^\top,$$

where

$$\frac{\partial H}{\partial \boldsymbol{\gamma}^\top} \quad = \quad \frac{\partial H}{\partial I}\frac{\partial I}{\partial \boldsymbol{\gamma}^\top} + \frac{\partial H}{\partial t^+_1}\frac{\partial t^+_1}{\partial \boldsymbol{\gamma}^\top} + \frac{\partial H}{\partial t^+_2}\frac{\partial t^+_2}{\partial \boldsymbol{\gamma}^\top},$$

and I denotes the identity function.

How to get $\widehat{\lambda}$ in the Box-Cox transformation approach

Consider the scaled Box-Cox transformation[14] with $W^{(\lambda)}_i = Y^{(\lambda)}_i / \tilde{Y}^{\lambda-1}$, where $\tilde{Y}$ is the geometric mean of all diagnostic test results, and $Y^{(\lambda)}_i$ is the Box-Cox transformed response. The linear mixed-effect model with new transformed responses becomes

$$\begin{aligned} W^{(\lambda)}_1 &= \alpha_{\star k_1} + z_1^\top \boldsymbol{\beta}_{\star 1} + \varepsilon_{\star 1}, \\ W^{(\lambda)}_2 &= \alpha_{\star k_2} + z_1^\top \boldsymbol{\beta}_{\star 2} + \varepsilon_{\star 2}, \\ W^{(\lambda)}_3 &= \alpha_{\star k_3} + z_1^\top \boldsymbol{\beta}_{\star 3} + \varepsilon_{\star 3}, \end{aligned} \tag{6.1}$$

with restricted log-likelihood

$$\begin{aligned}\ell_R^\star(\boldsymbol{\gamma}_\star;\lambda) &= -\frac{1}{2}\sum_{k=1}^{c}\left(\mathbf{W}_k^{(\lambda)}-\mathbf{Z}_k\widehat{\boldsymbol{\beta}}_{\star\lambda}(\boldsymbol{\theta}_\star)\right)^\top\boldsymbol{\Sigma}_{\star k}^{-1}\left(\mathbf{W}_k^{(\lambda)}-\mathbf{Z}_k\widehat{\boldsymbol{\beta}}_{\star\lambda}(\boldsymbol{\theta}_\star)\right)-\frac{1}{2}\sum_{k=1}^{c}\log|\boldsymbol{\Sigma}_{\star k}|\\ &\quad-\frac{1}{2}\log\left|\sum_{k=1}^{c}\mathbf{Z}_k^\top\boldsymbol{\Sigma}_{\star k}^{-1}\mathbf{Z}_k\right|,\end{aligned}$$

where $\mathbf{W}_k^{(\lambda)}$ is the n_k-vector of the scaled Box-Cox transformed responses within the cluster k-th, $\boldsymbol{\Sigma}_{\star k}=\sigma_{\star c}^2\mathbf{V}_k\mathbf{V}_k^\top+diag\{\sigma_{\star 1}^2,\ldots,\sigma_{\star 1}^2;\sigma_{\star 2}^2,\ldots,\sigma_{\star 2}^2;\sigma_{\star 3}^2,\ldots,\sigma_{\star 3}^2\}_{n_k}$ with $\mathbf{V}_k$ as $\mathbf{1}_{n_k}$, and

$$\widehat{\boldsymbol{\beta}}_{\star\lambda}(\boldsymbol{\theta}_\star)\equiv\widehat{\boldsymbol{\beta}}_\star(\boldsymbol{\theta}_\star,\lambda)=\left(\sum_{k=1}^{c}\mathbf{Z}_k^\top\boldsymbol{\Sigma}_{\star k}^{-1}\mathbf{Z}_k\right)^{-1}\sum_{k=1}^{c}\mathbf{Z}_k^\top\boldsymbol{\Sigma}_{\star k}^{-1}\mathbf{W}_k^{(\lambda)}.$$

Then, we find $\widehat{\lambda}$ based on a grid search, i.e., as

$$\widehat{\lambda}=\underset{\lambda\in[-2,2]}{\arg\max}\,\ell_R^\star(\widehat{\boldsymbol{\gamma}}_{\star\lambda};\lambda),$$

where $\widehat{\boldsymbol{\gamma}}_{\star\lambda}$ maximizes $\ell_R^\star(\boldsymbol{\gamma}_\star;\lambda)$, for a fixed value of λ. The value $\widehat{\lambda}$ obtained from the scaled model (6.1) is the same value that one would get from the original transformation model (3.9) (see [14] and [22]).