Statistical Methods in Medical Research
XX(X):2–34

DOI: 10.1177/ToBeAssigned
ijr.sagepub.com/

A distribution-free smoothed combination method to improve discrimination accuracy in multi-category classification

Raju Maiti[1], Jialiang Li[2], Priyam Das[3], Xueqing Liu[4], Lei Feng[5], Derek J Hausenloy[6, 7, 8, 9, 10], Bibhas Chakraborty[2,4,11]

Abstract

Results from multiple diagnostic tests are combined in many ways to improve the overall diagnostic accuracy. For binary classification, maximization of the empirical estimate of the area under the receiver operating characteristic (ROC) curve has widely been used to produce an optimal linear combination of multiple biomarkers. However, in the presence of large number of biomarkers, this method proves to be computationally expensive and difficult to implement since it involves maximization of a discontinuous, non-smooth function for which gradient-based methods cannot be used directly. Complexity of this problem further increases when the classification problem becomes multi-category. In this article, we develop a linear combination method that maximizes a smooth approximation of the empirical Hyper-volume Under Manifolds (HUM) for multi-category outcome. We approximate HUM by replacing the indicator function with the sigmoid function and normal cumulative distribution function (CDF). With such smooth approximations, efficient gradient-based algorithms are employed to obtain better solutions with less computing time. We show that under some regularity conditions, the proposed method yields consistent estimates of the coefficient parameters. We derive the asymptotic normality of the coefficient estimates. A simulation study is performed to study the effectiveness of our proposed method as compared to other existing methods. The method is illustrated using two real medical data sets.

Keywords

Hyper-volume Under the Manifolds (HUM); Volume under the surface (VUS); Multi-category classification; Sigmoid approximation; Acute kidney injury; Alzheimer disease

Prepared using sagej.cls *[Version: 2015/06/09 v1.01]*

1. Introduction

Statistical classification methods are widely used in various fields of science such as economics, computer science, meteorology, medicine, and many others. Specifically, in medicine, multiple diagnostic tests are combined in many ways to discriminate diseased individuals from the non-diseased. Over the last two decades, many research articles recommended combining multiple diagnostic test results in order to increase the overall diagnostic accuracy. Common approaches to combine multiple test results include the logistic regression (LR), the linear discriminant analysis (LDA) and other model-based approaches. Some authors ((1), (2), (3)) directly focused on the maximization of the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) to combine multiple test results. In brief, ROC curve for a classifier $h(\mathbf{X})$ is drawn through the set of points $\{(\text{TPR}(c), \text{FPR}(c)), -\infty < c < \infty\}$, where the true positive rate (TPR) and false positive rate (FPR) are defined as $\text{TPR}(c) = P(h(\mathbf{X}_i) > c | i\text{-th individual is diseased})$, and $\text{FPR}(c) = P(h(\mathbf{X}_j) > c |\ j\text{-th individual is non-diseased})$, respectively. An ROC curve is often summarized by the area under the ROC curve (AUC) to estimate the discrimination accuracy of a classifier. However, to the best of our knowledge, there are limited developments for finding an optimal linear combination of biomarkers in case of multi-category disease classification. Our objective here is to develop a classifier for multi-category outcome based on the multi-category version of AUC which is commonly known as the Hyper-volume Under the ROC Manifold (HUM).

For binary classification, earlier works considered maximizing various non-parametric estimates of AUC to obtain the best linear combination of the features (see (1), (2), (3), (4), (5) and few others). In particular, (3) proposed to maximize an empirical estimate of AUC in the form of a Mann-Whitney U-statistic for obtaining an optimal linear combination. However, maximization of the empirical AUC remains computationally challenging since the objective function is discontinuous and non-differentiable. To reduce the computational

[1]Economic Research Unit, Indian Statistical Institute Kolkata
[2]Department of Statistics and Applied Probability, National University of Singapore, Singapore
[3]Department of Biomedical Informatics, Harvard Medical School, USA
[4]Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore
[5]Department of Psychological Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[6]Cardiovascular and Metabolic Disorders Program, Duke-NUS Medical School, Singapore
[7]National Heart Research Institute Singapore, National Heart Centre, Singapore
[8]Yong Loo Lin School of Medicine, National University Singapore, Singapore
[9]The Hatter Cardiovascular Institute, University College London, London, UK
[10]Cardiovascular Research Center, College of Medical and Health Sciences, Asia University, Taiwan
[11]Department of Biostatistics and Bioinformatics, Duke University, USA

Corresponding author:
Economic Research Unit, Indian Statistical Institute Kolkata
Email: rajumaiti@gmail.com

Prepared using sagej.cls

burden, (4) proposed to maximize a smooth approximation of the empirical AUC using the sigmoid function. (5) proposed the Min-Max method where only two biomarkers with minimum and maximum values were combined by maximizing the empirical AUC. Thus, irrespective of the number of biomarkers, Min-Max method estimates only one coefficient parameter which makes it computationally much less challenging.

When a disease outcome is multi-category, the Hyper-volume Under the ROC Manifold (HUM) is often used as a diagnostic accuracy measure, a multi-category extension of AUC ((6)). Like the binary classification, here also one can directly maximize the HUM to combine multiple biomarkers. For a three-category disease outcome, HUM is known as the Volume under the ROC Surface (VUS), and has been used in a few real applications ((7), (8), (9)). (10) maximized the empirical estimate of VUS to combine multiple biomarkers. Due to non-differentiability of the objective function, maximization of empirical VUS requires derivative-free optimization methods which are computationally expensive, especially when the number of biomarkers is large. To overcome this problem, (11) used a penalized and scaled stochastic distance method assuming that biomarkers were normality distributed. Such method is computationally less challenging. However, violation of the normality assumption of biomarkers may lead to poor estimation performance. (12) constructed upper and lower bounds of the HUM using Fréchet inequality and maximized these bounds to combine multiple biomarkers. They showed that these bounds were functions of AUCs of all possible pairwise adjacent categories, and hence computationally the method is less challenging especially for large number of disease categories. However, such approximations do not perform well for small sample size and/or non-normal distributions (as is noticed in our simulation study). To study the discrimination power of a single biomarker or multiple biomarkers, one may adopt in-built R packages `HUM` (see (13)) and `mcca` (see (14)).

In this article, we propose to maximize a distribution-free Smooth approximation of the empirical HUM (SHUM) to combine multiple biomarkers in an effective way. In particular, the sigmoid function and the normal cumulative distribution functions (CDF) are used to approximate the non-differentiable indicator functions embedded in the definition of HUM. We show that the proposed method yields consistent estimates of the optimal coefficients and they are asymptotically normal. A major advantage with the proposed method stems from the fact that SHUM is a continuous and differentiable function; this allows one to adopt a variety of gradient-based optimization algorithms. Maximizing the empirical HUM with derivative-free optimization techniques, such as Nelder-Mead simplex method, genetic algorithm (GA), and simulated annealing (SA), are computationally expensive. However, gradient-based optimization techniques like Newton-Raphson and Quasi-Newton methods can be applied to maximize the SHUM function; these nonlinear solvers are much more stable with nice convergence properties. In addition to the theoretical developments, we also carry out extensive simulations to facilitate comparison with other existing methods, e.g., the min-max method

((5)), the lower and upper bound methods ((12)), the empirical method ((10)), and the parametric method with normal distribution ((15)).

As illustrative applications, we first consider data from the **E**ffect of **R**emote **I**schemic Preconditioning on **C**linical Outcomes in Patient Undergoing **C**oronary **A**rtery Bypass Graft Surgery (ERICCA) trial where a group of patients participated in a cardiovascular surgery and were followed for one year after the surgery ((16)). During the study period, some patients developed the Acute Kidney Injury (AKI) disease which was recorded as a multi-category ordinal outcome with four severity levels. In another application, we analyze data on Alzheimer's disease from the Alzheimer's Disease Research Center (ADRC) at the University of Washington. In this data set, the patients were divided into 3 groups based on the level of disease severity and data on 14 biomarkers were collected. For both the datasets, we apply our proposed method to combine the biomarkers and compare the results with existing methods.

The rest of the article is organized as follows. In Section 2, HUM and SHUM are defined along with discussion on the large sample properties of the estimated combination coefficients. In Section 3, existing methods are summarized as an overview. In Section 4, we provide a discussion on computational issues. In Section 5, we present results from the simulation studies. Section 6 describes the findings from two real data analyses. Section 7 contains discussion and concluding remarks. All the proofs of theoretical results appear in the Appendix. **R** codes are publicly available at *https://github.com/rajumaiti/SHUM.*

2. Proposed Estimators and Their Theoretical Properties

In this section, we introduce the HUM, empirical HUM and SHUM methods for combining multiple biomarkers to improve the multi-category discrimination accuracy.

Hyper-volume Under Manifolds (HUM)

Consider a study where there are M classes of the outcome variable which are assumed to be ordered in nature. We provide some practical suggestion later for unordered classes. Suppose $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_M$ are d-dimensional random selected vectors representing the values of d biomarkers for M outcome categories where $\mathbf{X}_j = (X_{j1}, X_{j2}, \cdots, X_{jd})^T$ and X_{jk} denotes the value of the k-th biomarker from the j-th category, $k = 1, 2, \cdots, d$ and $j = 1, 2, \cdots, M$. Suppose $\mathbf{X}_j$ follows a multivariate continuous distribution F_j. Consider a linear combination of these biomarkers as

$$\boldsymbol{\beta}^T \mathbf{X}_j = \sum_{k=1}^{d} \beta_k X_{jk},\ j = 1, 2, \cdots, M,$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_d)^T$ is a d-dimensional vector of parameters. Under the assumption that larger values of $\boldsymbol{\beta}^T \mathbf{X}$ corresponds to more severe disease categories, a discrimination accuracy measure can be defined by the following

probability

$$D(\boldsymbol{\beta}) = P(\boldsymbol{\beta}^T \mathbf{X}_M > \boldsymbol{\beta}^T \mathbf{X}_{(M-1)} > \cdots > \boldsymbol{\beta}^T \mathbf{X}_1),$$

which is known as hyper-volume under the ROC manifold (HUM) ((7), (6)). For multi-category ordinal outcomes, HUM can be considered as an extension of the AUC which is widely used in binary discrimination accuracy studies. Here our objective is to find the best possible value of $\boldsymbol{\beta}$ for which $D(\boldsymbol{\beta})$ is maximum. Ideally, if there exists a $\boldsymbol{\beta}$ for which $D(\boldsymbol{\beta}) = 1$, then such a linear combination perfectly separates all the classes. Letting $\boldsymbol{\beta}_0$ denote the optimal coefficient parameter that maximizes $D(\boldsymbol{\beta})$ over a restricted parametric space $B = \{\boldsymbol{\beta} \in \mathbb{R}^d : \beta_d = 1\}$, then we can write

$$\boldsymbol{\beta}_0 = \arg\max_{\boldsymbol{\beta} \in B} D(\boldsymbol{\beta}).$$

Note that we assume the d-th component β_d of the coefficient vector $\boldsymbol{\beta}$ to be 1 in order to avoid the identifiability problem. Denote $\boldsymbol{\theta} = (\beta_1, \beta_2, \cdots, \beta_{d-1})^T$ to be the first $d-1$ components of $\boldsymbol{\beta}$ which are free to take any value in the $d-1$ dimensional Euclidean space. Hereafter, for the simplicity of notation, we use $\boldsymbol{\beta}$ in place of $\boldsymbol{\beta}(\boldsymbol{\theta}) = (\boldsymbol{\theta}^T, 1)^T$. If the biomarkers are non-informative in predicting the outcome category then $D(\boldsymbol{\beta})$ will be close to $\frac{1}{M!}$ which is the probability of a random sorting. Under the assumption that $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_M$ are generated from multivariate normal distribution and some mild regularity conditions, an unique optimal solution of $\boldsymbol{\beta}$, which has nice closed form expression, can be derived ((1)). However, in general for non-normal data, there does not exist any closed form expression of $\boldsymbol{\beta}_0$ and hence a numerical optimizer should be employed to know the approximate true value with a very large sample size and a large Monte Carlo size which is a standard practice often used in many statistical problems to know the true parameter value.

Empirical Hyper-volume Under Manifolds (EHUM)

Now let us consider the problem of estimating $\boldsymbol{\beta}_0$ given an empirical sample. Let $\{\mathbf{X}_{ji_j};\ i_j = 1, 2, \cdots, n_j,\ j = 1, 2, \cdots, M\}$ be a sample of size $n = \sum_{j=1}^{M} n_j$ observations where $j = 1, \ldots, M$ denote outcome categories and $i_j = 1, 2, \cdots, n_j$ denote the samples in the j-th category. Then, for a fixed $\boldsymbol{\beta}$, the empirical HUM can be written as

$$\begin{aligned} D_E(\boldsymbol{\beta}) &= \frac{1}{n_1 n_2 \cdots n_M} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \cdots \sum_{i_M=1}^{n_M} I(\boldsymbol{\beta}^T \mathbf{X}_{Mi_M} > \boldsymbol{\beta}^T \mathbf{X}_{(M-1)i_{(M-1)}} > \cdots > \boldsymbol{\beta}^T \mathbf{X}_{1i_1}) \\ &= \frac{1}{\prod_{j=1}^{M} n_j} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \cdots \sum_{i_M=1}^{n_M} I(\boldsymbol{\beta}^T \mathbf{X}_{Mi_M} > \boldsymbol{\beta}^T \mathbf{X}_{(M-1)i_{(M-1)}}) \cdots I(\boldsymbol{\beta}^T \mathbf{X}_{2i_2} > \boldsymbol{\beta}^T \mathbf{X}_{1i_1}) \end{aligned} \tag{1}$$

where $I(\cdot)$ denotes the indicator function. When the sample size n is large, the empirical estimator $D_E(\boldsymbol{\beta})$ is a very close to the original function $D(\boldsymbol{\beta})$.

Therefore, an optimal coefficient vector can be estimated by maximizing the empirical estimate $D_E(\boldsymbol{\beta})$ which can be written as

$$\widehat{\boldsymbol{\beta}}_E = \arg\max_{\boldsymbol{\beta}\in B} D_E(\boldsymbol{\beta}).$$

When the number of disease categories is 2 (i.e., $M = 2$), the empirical HUM reduces to the empirical estimate of AUC (see (3)) given by

$$D_E(\boldsymbol{\beta}) = \frac{1}{n_1 n_2} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} I(\boldsymbol{\beta}^T \mathbf{X}_{2i_2} > \boldsymbol{\beta}^T \mathbf{X}_{1i_1}),$$

and when $M = 3$, it reduces to the empirical VUS (see (10)) given by

$$D_E(\boldsymbol{\beta}) = \frac{1}{n_1 n_2 n_3} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \sum_{i_3=1}^{n_3} I(\boldsymbol{\beta}^T \mathbf{X}_{3i_3} > \boldsymbol{\beta}^T \mathbf{X}_{2i_2} > \boldsymbol{\beta}^T \mathbf{X}_{1i_1}). \tag{2}$$

Under some regularity conditions, (10) established the consistency and asymptotic normality of $\widehat{\boldsymbol{\beta}}_E$ for three-category outcomes. Following their argument, consistency and asymptotic normality of $\widehat{\boldsymbol{\beta}}_E$ for more than three categories can easily be established. However, upon close examination we notice that $D_E(\boldsymbol{\beta})$ is discontinuous and not differentiable with respect to $\boldsymbol{\beta}$, and hence faster gradient-based algorithms are not useful to this optimization problem. On the other hand, although derivative-free algorithms can be used for smaller number of categories, say $M = 2$ or 3, such algorithms become computationally prohibitive and unstable as the number of categories increases. To overcome this problem, in the next section, we propose a new method based on a smooth approximation of the empirical HUM.

Smooth Approximations of empirical HUM (SHUM)

In order to alleviate the computational burden of maximizing the empirical HUM $D_E(\boldsymbol{\beta})$, we propose to maximize a class of smooth approximations of the empirical HUM as an alternative approach. The basic idea is to approximate the non-differentiable indicator function $I(x > 0)$ by a smooth function. We consider a class of all continuous distribution functions $g(x)$ with support space $(-\infty, \infty)$, satisfying $g(x) + g(-x) = 1$ and $g''(x)$ is continuous. Replacing all the indicator functions with this $g(x)$ functions in the $D_E(\boldsymbol{\beta})$ makes the approximate objective function solvable with the gradient-based optimization algorithms such as the Newton-Raphson method and the Quasi-Newton method. In this paper, we consider two smooth candidates from the above class to solve the computational issue. The first one is the sigmoid function $s(x) = \frac{1}{1+\exp(-x)}$, and the second one is the standard normal CDF denoted by $\Phi(x) = P(\chi \leq x)$ where χ follows a normal distribution with mean 0 and variance 1.

Note that when x is close to 0, the absolute difference between $s(x)$ and $I(x)$ is the highest and as x goes away from 0, $s(x)$ gets closer to $I(x)$. This is also true for $\Phi(x)$. However, we can improve these approximations $s(x)$

and $\Phi(x)$ by introducing a tuning parameter λ into these functions which are given as follows: $s_n(x) = s(\frac{x}{\lambda}) = \frac{1}{1+\exp(-x/\lambda)}$ and $\Phi_n(x) = \Phi(x/\lambda)$ where λ satisfies $\lim_{n\to\infty} \lambda = 0$. The choice of the tuning parameter λ is very crucial in approximating the indicator function $I(x > 0)$ and hence the empirical HUM.

When λ is close to 0, the proposed SHUM estimator behaves similarly to the empirical HUM with a very large value of derivative across a very small interval around zero. This induces a greater variability on the resulting estimators. On the other hand, if λ is chosen to be one, it suffers from biased approximation. Therefore, we need to choose an optimal λ between 0 and 1 to strike a balance between the bias and the variance issues. To illustrate the role of λ, a graphical representation is displayed in Figure 2 where we consider a few selected values of λ_n. We can see that as λ decreases to zero the approximation becomes closer to the indicator function $I(x > 0)$. As a rule of thumb, for binary classification problem (17) and (4) recommended the value of λ for which $|\boldsymbol{\beta}^T(\mathbf{X}_{1i_1} - \mathbf{X}_{2i_2})/\lambda| > 5$ for most of the pairs $(i_1,\ i_2)$. In our case, λ should satisfy

$$|\boldsymbol{\beta}^T(\mathbf{X}_{1i_1} - \mathbf{X}_{2i_2})/\lambda||\boldsymbol{\beta}^T(\mathbf{X}_{2i_2} - \mathbf{X}_{3i_3})/\lambda|\cdots|\boldsymbol{\beta}^T(\mathbf{X}_{(M-1)i_{(M-1)}} - \mathbf{X}_{Mi_M})/\lambda| > 5$$

for most of the pairs $(i_1, i_2, \cdots, i_M)$. In this regard, a possible choice for λ is $\frac{1}{\sqrt{n(M-1)}}$, a general extension of the choice of $\lambda = \frac{1}{\sqrt{n}}$ for binary classification.

We performed some simulation experiments to empirically find the best possible value of the tuning parameter λ_n in the definition of SHUM given in equation (**??**). We simulated samples from the multivariate normal distributions with $M = 3$ and $d = 4$. We explored the samples of size $n = (20, 20, 20)$ and $n = (50, 50, 50)$. We calculated the variance and bias for varying values of λ_n and presented the results in Figure 1. As we can see, for total sample 60, variance increases as λ_n increases and it gets stable after $\frac{1}{\lambda_n} = 10 \approx \sqrt{(3-1)\times 60}$ and for total sample size 150, the varianace is stable after $\frac{1}{\lambda_n} = 17 \approx \sqrt{(3-1)\times 150}$. In case of bias, it decreases as λ_n increases and it gets stable after $\frac{1}{\lambda_n} = 10$ for total sample size 60, and $\frac{1}{\lambda_n} = 17$ for total sample size 150.

Although $D_E(\boldsymbol{\beta})$ can be approximated using any of the smoothed functions $s_n(x)$ and $\Phi_n(x)$, hereafter we only derive the results using the sigmoid function $s_n(x)$ to save the space. The proposed sigmoid function based approximation of $D_E(\boldsymbol{\beta})$ is given by

$$D_{s_n}(\boldsymbol{\beta}) = \frac{1}{n_1 n_2 \cdots n_M} \sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\cdots\sum_{i_M=1}^{n_M} s_n(\boldsymbol{\beta}^T(\mathbf{X}_{Mi_M} - \mathbf{X}_{(M-1)i_{(M-1)}}))\cdots s_n(\boldsymbol{\beta}^T(\mathbf{X}_{2i_2} - \mathbf{X}_{1i_1})), \tag{3}$$

where all the indicator functions are replaced by the sigmoid functions. We propose to maximize $D_{s_n}(\boldsymbol{\beta})$ in order to estimate the optimal coefficient vector. The optimal coefficient vector is then given by

$$\widehat{\boldsymbol{\beta}}_{s_n} = \arg\max_{\boldsymbol{\beta}\in B} D_{s_n}(\boldsymbol{\beta}).$$

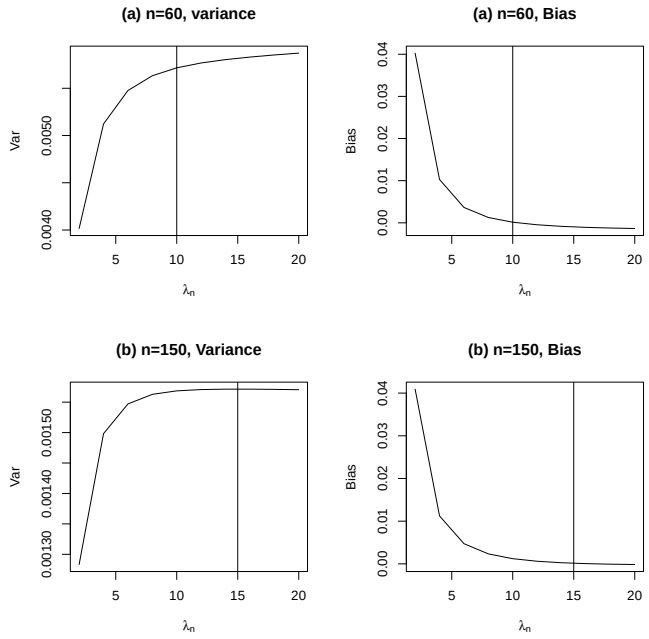

Figure 1. Choice of λ_n using bias-variance trade-off.

We denote the optimal coefficient estimate obtained using the sigmoid smooth approximation of the empirical HUM (SSHUM) as $\widehat{\boldsymbol{\beta}}_{s_n}$ and using the normal smooth approximation of the empirical HUM (NSHUM) by $\widehat{\boldsymbol{\beta}}_{\Phi_n}$ for the simulation and data analysis.

Consistency and Asymptotic Normality of SSHUM

Under some regularity conditions, we establish the consistency and asymptotic normality of $\widehat{\boldsymbol{\beta}}_{s_n}$. These conditions are listed as follows.

A1. The support space of $\mathbf{X}_{ji_j}$ is not contained in any proper linear subspace of $\mathbb{R}^d$.
A2. There exist at least one component of $\mathbf{X}_{ji_j}$ that has positive density everywhere conditional on the other components, almost surely.
A3. The true parameter value $\boldsymbol{\beta}_0$ is an interior point of B which is a compact subset of $\mathbb{R}^d$.

Theorem 1. Consistency. *Suppose that assumptions (A1)-(A3) hold, then*

$$\widehat{\boldsymbol{\beta}}_{s_n} \xrightarrow{p} \boldsymbol{\beta}_0$$

as $n \to \infty$, where "$\xrightarrow{p}$" denotes convergence in probability.

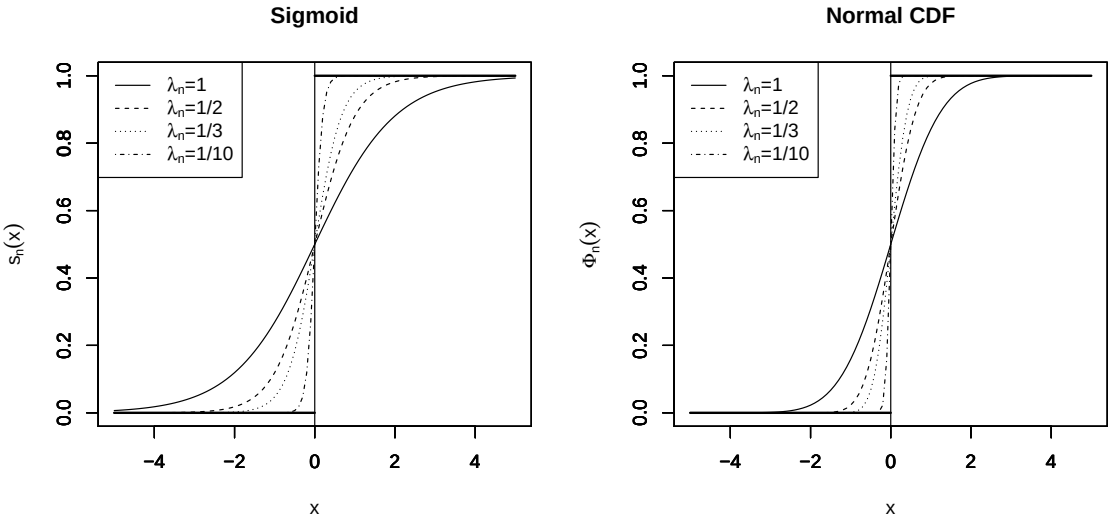

Figure 2. Sigmoid and normal CDF functions for different choices of tuning parameter λ_n

The detailed proof of Theorem 1 is provided in Section A2 in Appendix.

In order to prove the asymptotic normality of $\widehat{\boldsymbol{\beta}}_{s_n}$, we assume an additional set of regularity conditions. Denote $\Psi(\mathbf{X}_{1i_1}, \mathbf{X}_{2i_2}, \mathbf{X}_{Mi_M}; \boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \left[s_n(\boldsymbol{\beta}^T(\mathbf{X}_{Mi_M} - \mathbf{X}_{(M-1)i_{(M-1)}})) \cdots s_n(\boldsymbol{\beta}^T(\mathbf{X}_{2i_2} - \mathbf{X}_{1i_1})) \right]$. Then assume the following:

A4. $\mathbf{A}(\boldsymbol{\beta}_0) = E\left(-\frac{\partial}{\partial \boldsymbol{\theta}^T} \Psi(\mathbf{X}_{1i_1}, \mathbf{X}_{2i_2}, \mathbf{X}_{Mi_M}; \boldsymbol{\beta}_0) \right) < \infty$ and is invertible.

A5. $\tilde{\Psi}_{m1}(\mathbf{X}_{m1}; \boldsymbol{\beta}_0) = E\left(\frac{\partial}{\partial \boldsymbol{\theta}} \Psi(\mathbf{X}_{11}, \mathbf{X}_{21}, \mathbf{X}_{M1}; \boldsymbol{\beta}_0) | \mathbf{X}_{m1} \right)$ has a finite variance-covariance matrix, i.e., $\Sigma_{\psi_m} = Var(\tilde{\Psi}_{m1}(\mathbf{X}_{m1}; \boldsymbol{\beta}_0)) < \infty$ for all $m = 1, 2, \cdots, M$.

A6. $\lim_{n\to\infty} \frac{n}{n_m} = \rho_m^2 < \infty$ for all $1 \le m \le M$.

Assumptions (A4)-(A6) ensure that the asymptotic variance exits and is finite.

Theorem 2. Asymptotic normality. *Suppose that the regularity conditions (A1)-(A6) hold, then*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{s_n} - \boldsymbol{\beta}_0) \xrightarrow{D} (W^T, 0)^T$$

as $n \longrightarrow \infty$ *where* "$\xrightarrow{D}$" *denotes convergence in distribution and* W *is a* $(d-1)$*-variate normal distribution* $N(\mathbf{0}, \mathbf{A}^{-1}(\boldsymbol{\beta}_0)\mathbf{B}(\boldsymbol{\beta}_0)\{\mathbf{A}^{-1}(\boldsymbol{\beta}_0)\}^T)$*, where*

$$\mathbf{B}(\boldsymbol{\beta}_0) \quad = \quad \sum_{m=1}^{M} \rho_m^2 \Sigma_{\psi_m}.$$

The proof is provided in Section A3 in Appendix.

Computation of variance of $\widehat{\boldsymbol{\beta}}_{s_n}$ using the asymptotic variance formula given in Theorem 2 is very tedious and challenging, especially because of the complicated nature of the smoothing function s_n and its first and second derivatives. Furthermore, it is observed that the U-statistic based asymptotic variance formula are not generally reliable for small sample size (see (6)). In such cases, bootstrap technique is commonly employed to compute the variances of the coefficient estimators $\boldsymbol{\beta}_{s_n}$.

3. Review of Competing Methods

In this section, we provide a brief summary of some existing combining methods which are based on direct maximization of different versions of AUC in the multi-category classification problem. In the simulation and data analysis sections, we will compare our proposed method with these existing methods.

Parametric Method with Normality Assumption (Parametric)

(15) proposed to maximize the HUM $D(\boldsymbol{\beta})$ under the assumption that biomarker vectors $\boldsymbol{X}_j$ from the j-th category follow multivariate normal distribution with mean vector $\boldsymbol{\mu}_j$ and variance-covariance matrix Σ_j, $j = 1, 2, \cdots, M$. Note that under muultivariate normal distributions any linear combination of biomarker vector $\boldsymbol{X}_j$, denoted by $V_j = \boldsymbol{\beta}^T \boldsymbol{X}_j$, follows a univariate normal distribution with mean $\boldsymbol{\beta}^T \boldsymbol{\mu}_j$ and variance $\boldsymbol{\beta}^T \Sigma_j \boldsymbol{\beta}$, i.e., $V_j \sim N(\boldsymbol{\beta}^T \boldsymbol{\mu}_j, \boldsymbol{\beta}^T \Sigma_j \boldsymbol{\beta})$, $j = 1, 2, \cdots, M$. Let ϕ and Φ denote the density function and the cumulative distribution function of the standard normal distribution $N(0, 1)$. Then, for $M = 3$, the HUM $D(\boldsymbol{\beta})$ can be written as

$$D_N(\boldsymbol{\beta}) \quad = \quad \int_{-\infty}^{\infty} \Phi\left(\frac{\sqrt{\boldsymbol{\beta}^T \Sigma_2 \boldsymbol{\beta}}}{\sqrt{\boldsymbol{\beta}^T \Sigma_1 \boldsymbol{\beta}}} u + \frac{\boldsymbol{\beta}^T (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)}{\sqrt{\boldsymbol{\beta}^T \Sigma_1 \boldsymbol{\beta}}} \right) \Phi\left(-\frac{\sqrt{\boldsymbol{\beta}^T \Sigma_2 \boldsymbol{\beta}}}{\sqrt{\boldsymbol{\beta}^T \Sigma_3 \boldsymbol{\beta}}} u + \frac{\boldsymbol{\beta}^T (\boldsymbol{\mu}_3 - \boldsymbol{\mu}_2)}{\sqrt{\boldsymbol{\beta}^T \Sigma_3 \boldsymbol{\beta}}} \right) \phi(u)\, du. \tag{4}$$

Maximizing $D_N(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, we can obtain the optimal coefficient estimates as

$$\widehat{\boldsymbol{\beta}}_N = \arg\max_{\boldsymbol{\beta} \in B} D_N(\boldsymbol{\beta}).$$

Following the results of (1), it can be shown that if $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_M$ are multivariate normally distributed with mean vectors $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\cdots$, $\boldsymbol{\mu}_M$, respectively satisfying

$$\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 = \boldsymbol{\mu}_3 - \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_M - \boldsymbol{\mu}_{M-1} = \boldsymbol{\delta}, \tag{5}$$

and common variance-covariance matrix Σ, then the optimal coefficient parameters $\widehat{\boldsymbol{\beta}}_N$ will be proportional to $\Sigma^{-1}\boldsymbol{\delta}$, i.e., $\widehat{\boldsymbol{\beta}}_N \propto \Sigma^{-1}\boldsymbol{\delta}$. For $M = 3$, this result is discussed in (10).

A major advantage of using normality assumption is that it is computationally less challenging, especially when (5) holds true. When Σ and δ are unknown, one needs to estimate those parameters from the data and plug-in them into the above formula of $\widehat{\boldsymbol{\beta}}_N$. However, the main limitation with this method is that it is fully parametric as it fully depends on the normality assumption. Violation of normality of the data may result in poor estimate of $\boldsymbol{\beta}_0$.

Min-Max Method (Min-Max)

The Min-Max (MM) method is a more simplified non-parametric approach to combine the multiple biomarkers. It was originally proposed by (5) in the context of binary outcome. Instead of considering all the biomarkers, this method considers the empirical AUC based on the linear combination of two extreme biomarkers for each subject in the study. In this paper, to facilitate a comparative study, we define the empirical HUM based on the combination of the minimum and maximum biomarkers for each subject.

Let $X_{ji_j,max} = \max_{1\le k\le d} X_{ji_j,k}$ and $X_{ji_j,min} = \min_{1\le k\le d} X_{ji_j,k}$ and define the linear combination of these two extreme observations as $V_{ji_j} = \beta_{max}X_{ji_j,max} + \beta_{min}X_{ji_j,min}, \quad i = 1, 2, \cdots, n_j, \quad j = 1, 2, \cdots, M$. Then the objective function to be maximized to obtain the optimal coefficient vector is given by

$$D_{MM}(\boldsymbol{\beta}) \quad = \quad \frac{1}{\prod_{j=1}^{M} n_j} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \cdots \sum_{i_M=1}^{n_M} I(V_{Mi_M} > V_{(M-1)i_{M-1}} > \cdots > V_{1i_1}) \quad (6)$$

The optimal coefficient estimates by maximizing the above quantity can be written as

$$\widehat{\boldsymbol{\beta}}_{MM} = \arg\max_{\boldsymbol{\beta}\in B} D_{MM}(\boldsymbol{\beta}).$$

A major advantage with this method is that it involves the optimization of a single parameter as opposed to other competing methods that consider several parameters depending on the number of biomarkers, and hence it is computationally very efficient. Furthermore, it need not assume any distributional assumption of the data and hence is more robust against the parametric methods. So far, the method has been studied only for binary disease outcome, in which case the method can achieve higher sensitivity over a certain range of specificity. In other words, when one is interested in partial AUC, the method works better. However, the main limitation with this method is that a major portion of the information on the biomarkers are not utilized since only maximum and minimum biomarker values are used.

Upper and Lower Bound Approach using Fréchet inequality (Fréchet)

To reduce computational burden of the maximization of the empirical HUM, (12) proposed to maximize the upper and lower bounds of the HUM which are given as follows

$$\max\{0, (M-1)P_A(\boldsymbol{\beta}) - (M-2)\} \leq D(\boldsymbol{\beta}) \leq P_M(\boldsymbol{\beta}),$$

where $P_A(\boldsymbol{\beta})$ and $P_M(\boldsymbol{\beta})$ are defined as follows

$$P_A(\boldsymbol{\beta}) = \sum_{j=1}^{M-1} P(\boldsymbol{\beta}^T\mathbf{X}_{j+1} > \boldsymbol{\beta}^T\mathbf{X}_j)/(M-1),$$

and

$$P_M(\boldsymbol{\beta}) = \min_{1\leq j\leq M-1} P(\boldsymbol{\beta}^T\mathbf{X}_{j+1} > \boldsymbol{\beta}^T\mathbf{X}_j).$$

For example, by maximizing the upper bound $P_M(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, we obtain $\widehat{\boldsymbol{\beta}}_{Frechet} = \arg\max_{\boldsymbol{\beta}\in B} P_M(\boldsymbol{\beta})$ which can be considered as an alternative optimal coefficient estimates. The above method is computationally efficient against the direct maximization of HUM as it only considers pairs from the adjacent categories and their corresponding AUCs. Hence the above method is computationally less time consuming than the direct maximization of the HUM especially when the number of outcome categories is relatively large. However, when the discrimination accuracy for the pairwise categories are not significant as compared to the overall discrimination, this method will fail to give good results in terms of overall discrimination accuracy.

4. Computational Considerations: Step-down Algorithm

Step-down algorithm was originally proposed by (3) to combine more than two biomarkers in case of binary disease outcomes. The main motivation of using step-down algorithm is its ability to optimize the elements of the $\boldsymbol{\beta}$ vector sequentially one at a time instead of optimizing them simultaneously. (11) formalized the step-down algorithm in the context of three-category diagnostic outcomes. Recently (12) used this algorithm to maximize upper or lower bound of HUM, namely P_M or P_A and obtained an optimal linear coefficient estimates. The algorithm to maximize a criteria function (e.g., EHUM) goes as follows:

Step 1. Compute the EHUM for each of the d biomarkers one at a time and arrange covariates in decreasing order with respect to the computed EHUM values such that $X_{(1)}$ and $X_{(d)}$ have the highest and the lowest individual EHUM values respectively.

Step 2. Choose the first two biomarkers with highest EHUM values and combine them as $V_2 = X_{(1)} + \lambda_2 X_{(2)}$.

Step 3. Maximize the EHUM for the combined biomarker V_2 w.r.t. λ_2 and obtain $\widehat{V}_2 = X_{(1)} + \widehat{\lambda}_2 X_{(2)}$.

Step 4. For $i = 3, \ldots, d$ construct $V_i = \widehat{V}_{i-1} + \lambda_i X_{(i)}$ and maximize V_i w.r.t. λ_i and obtain $\widehat{\lambda}_i$.

Thus at the end of step 4, the estimated optimal combination $\widehat{V}_d = X_{(1)} + \widehat{\lambda}_2 X_{(2)} + \cdots + \widehat{\lambda}_d X_{(d)}$ can be obtained. Note that the step-down algorithm can be employed to find optimal coefficient parameters by maximizing any function like EHUM, SHUM, P_M, etc. However, in general, the algorithm can be used to maximize any other function of multidimensional parameters. For computational advantages, the step-down algorithm has been used throughout the simulation studies and data analysis. We implement the numerical method using the in-built function `optim` in the `R` software freely available in `https://cran.r-project.org/`.

5. Simulation Study

In this section, we present results from simulation experiments to study the performance of the proposed methods compared to a few existing methods such as the Min-Max method, the Upper bound approach using Fréchet inequality, and the direct maximization of empirical HUM. All these existing methods are discussed in Section 3. Different simulation setups by varying the number of disease categories and number of biomarkers along with different multivariate distributions of biomarkers are explored.

Case 1: $d = 3, M = 3$

In this setup, three biomarkers (i.e., $d = 3$) and three disease categories (i.e., $M = 3$) are considered. Samples of sizes 60, 90 and 120 were used to perform this study. The biomarkers' values are generated from four different multivariate distributions. First three distributions are multivariate normal distributions with different covariance matrices, namely (1) independent, (2) exchangeable and (3) AR(1); and the fourth distribution is Weibull distribution which represents the family of skewed distributions.

Scenario 1 : For the i-th category, the values of the biomarkers are simulated from three dimensional normal distributions with mean vector $\boldsymbol{\mu}_i$, and common covariance matrix as identity $\Sigma = \mathbf{I}$, $i = 0, 1, 2$. We set the parameter values as $\boldsymbol{\mu}_0 = (0, 0, 0)^T$, $\boldsymbol{\mu}_1 = (1.0, 1.1, 1.2)^T$, and $\boldsymbol{\mu}_2 = (2.0, 2.2, 2.4)^T$ for categories $i = 0, 1, 2$, respectively. Since the correlation matrix is considered to be identity with normal distributions, the biomarkers are independent to each other.

Scenario 2 : In the second scenario, the mean vectors are same as in Scenario 1, however the covariance matrix $\Sigma = ((\sigma_{st}))$ is such that all the diagonal elements are 1, i.e, $\sigma_{ss} = 1$; and all the off-diagonal elements are 0.2, i.e., $\sigma_{st} = 0.2, s \neq t$; $s, t = 1, 2, 3$. This covariance matrix is an example of an exchangeable matrix. Since all the off-diagonal elements are non-zero and equal, therefore the biomarkers are correlated.

Scenario 3 : In the third scenario, the mean vectors are same as the previous scenarios. The covariance matrix has an AR(1) form, i.e., all the diagonal elements are 1; and the off-diagonal elements are set as $\sigma_{st} = 0.2^{|s-t|}$, $s \neq t$; $s, t = 1, 2, 3$. Here all the mutual correlations are non-zero but it fades as the distance between two biomarkers increases.

Scenario 4 : In the fourth scenario, values of the biomarkers are simulated from Weibull distribution. Specifically, the j-th biomarker from the i-th disease category follows a Weibull distribution with shape parameter k_j and scale parameter α_i and the probability density function is given by

$$f(x; k_j, \alpha_i) = \begin{cases} \frac{k_i}{\alpha_i}\left(\frac{x}{\alpha_i}\right)^{k_j-1} \exp(-(\frac{x}{\alpha_j})^{k_j}) & x > 0, \\ 0 & x \leq 0, \end{cases}$$

$i = 0, 1, 2$ and $j = 1, 2, 3$. Values of the shape parameter k and scale parameter λ are set as $(k_1, k_2, k_3) = (0.5, 1, 1.5)$ and $(\alpha_1, \alpha_2, \alpha_3) = (1, 2, 3)$, respectively. Here, we assume that biomarkers are independently distributed. This case corresponds to non-normal and skewed distribution.

For each fixed sample size, we simulate samples from each of the above scenarios and estimate the optimal coefficient vector $\boldsymbol{\beta}$ by maximizing the different versions of HUM. The process is repeated for 500 times. Based on that 500 results, we reported the mean and standard errors of HUM in Table 1, whereas the mean values for the coefficient vector are reported in Table 2.

Under the first three scenarios where biomarkers' values are generated from normal distributions, the SSHUM and NSHUM methods perform as good as the parametric method given in Section 3.1 and outperform the Fréchet bounds and Min-Max methods with respect to the discriminating accuracy. In Scenario 4, where biomarkers' values are non-normally distributed, the parametric method with normality assumption performs poorly compared to the proposed methods. However, there is no observable difference in the accuracy measure between the SSHUM and NSHUM methods suggesting that both the sigmoid and the normal CDF approximations perform equally well for non-normal distributions.

Case 2: $d = 4, M = 4$

In this setup, we considered four biomarkers (i.e., $d = 4$) and four disease categories (i.e., $M = 4$) to see how our proposed method performs for a larger number of disease categories. Samples of sizes 20, 30, and 40 were explored to perform this study. We consider lower sample size as compared to the $d = 3, M = 3$ case owing to the increased computational burden in the present case. The biomarkers are generated from three different multivariate normal distributions with different correlation matrices as in Case 1. The mean vectors for four disease classes are $\boldsymbol{\mu}_0 = (0, 0, 0, 0)^T$, $\boldsymbol{\mu}_1 = (0.1, 0.1, 0.1, 0.1)^T$, and $\boldsymbol{\mu}_2 = (0.2, 0.2, 0.2, 0.2)^T$ and $\boldsymbol{\mu}_3 = (0.3, 0.3, 0.3, 0.3)^T$, respectively.

Performance of the proposed SSHUM and NSHUM methods are compared with the existing methods, namely the empirical method ((10)), the Fréchet bounds method ((12)), and the Min-Max method ((5)). Since samples are only generated from multivariate normal distribution with different correlation structures, we do not include the parametric method in our comparison study. By default, the parametric method outperforms the others. Summarized results are presented in Table 3. As we can see from Table 3, the proposed SSHUM and

NSHUM outperforms the others in terms of the HUM values. Also it is worth noting that as the correlation between biomarkers gets weaker, the HUM value for all the methods increase.

Case 3: $d = 10, M = 4$

In this setup, we considered 10 biomarkers (i.e., $d = 10$) and four disease categories (i.e., $M = 4$). Samples of sizes 20, 30, and 40 were explored. The biomarkers are generated from three different multivariate normal distributions with different correlation matrices as in Case 1. The mean vectors for four disease classes are $\boldsymbol{\mu}_0 = (0, 0, \cdots, 0)^T$, $\boldsymbol{\mu}_1 = (0.1, 0.1, \cdots, 0.1)^T$, and $\boldsymbol{\mu}_2 = (0.2, 0.2, \cdots, 0.2)^T$ and $\boldsymbol{\mu}_3 = (0.3, 0.3, \cdots, 0.3)^T$, respectively.

As in case 3, the proposed SSHUM and NSHUM methods are compared with the empirical method ((10)), the Fréchet bounds method ((12)), and the Min-Max method ((5)). Table 4 summarize the simulation results. We observe that the proposed SSHUM and NSHUM methods outperform the others in terms of the HUM values. The Min-Max method has the worst performance. The results suggest that the proposed methods perform well even with a large number of biomarkers.

Table 1. Means and standard errors (in parenthesis) of obtained EHUM values at the optimal coefficient vector estimated using the methods: the Empirical method ((10)), the Fréchet bounds method ((12)), the parametric method ((15)), the Min-Max method, SSHUM and NSHUM for simulation Scenarios 1, 2, 3, 4 with sample sizes $(60, 60, 60), (90, 90, 90), (120, 120, 120)$, based on 1000 Monte Carlo replications.

(n_1, n_2, n_3)	Empirical	Min-Max	Parametric	Fréchet	SSHUM	NSHUM
			Scenario 1			
(60, 60, 60)	0.824 (0.032)	0.804 (0.035)	0.826 (0.032)	0.813 (0.034)	0.828 (0.033)	0.828 (0.033)
(90, 90, 90)	0.825 (0.026)	0.805 (0.028)	0.827 (0.027)	0.815 (0.026)	0.827 (0.026)	0.827 (0.026)
(120, 120, 120)	0.824 (0.022)	0.804 (0.023)	0.825 (0.022)	0.813 (0.022)	0.825 (0.022)	0.825 (0.022)
			Scenario 2			
(60, 60, 60)	0.747 (0.039)	0.734 (0.039)	0.752 (0.039)	0.744 (0.039)	0.754 (0.039)	0.754 (0.039)
(90, 90, 90)	0.748 (0.032)	0.735 (0.032)	0.750 (0.031)	0.744 (0.032)	0.752 (0.031)	0.752 (0.031)
(120, 120, 120)	0.749 (0.026)	0.736 (0.027)	0.751 (0.026)	0.745 (0.027)	0.752 (0.026)	0.752 (0.026)
			Scenario 3			
(60, 60, 60)	0.766 (0.037)	0.752 (0.038)	0.770 (0.037)	0.756 (0.039)	0.773 (0.037)	0.773 (0.037)
(90, 90, 90)	0.767 (0.030)	0.753 (0.031)	0.769 (0.031)	0.756 (0.031)	0.771 (0.030)	0.771 (0.030)
(120, 120, 120)	0.769 (0.026)	0.754 (0.026)	0.770 (0.026)	0.758 (0.026)	0.771 (0.026)	0.772 (0.026)
			Scenario 4			
(60, 60, 60)	0.452 (0.059)	0.412 (0.044)	0.436 (0.057)	0.391 (0.043)	0.521 (0.045)	0.521 (0.046)
(90, 90, 90)	0.474 (0.051)	0.412 (0.036)	0.425 (0.058)	0.391 (0.038)	0.515 (0.036)	0.515 (0.036)
(120, 120, 120)	0.484 (0.046)	0.411 (0.031)	0.420 (0.047)	0.392 (0.033)	0.512 (0.031)	0.512 (0.031)

Computation Time Evaluation

In the previous section it is noted that the proposed methods SSHUM and NSHUM outperforms other existing methods while not much difference in performance is noted between them. Also among existing methods, EVUS performs the best and closest to the proposed methods. Here we compare the computation times required for EVUS and SSHUM methods for one normal (Scenario 1) and one non-normal

Table 2. Means (biases and standard errors) of $(\beta_1, \beta_2)^T$ (based on 1000 replications) by different methods for Scenario 1. All the methods were maximized using Quasi-Newton method.

Sample size	$(\beta_1, \beta_2)^T$	Empirical	Parametric	Fréchet	SSHUM	NSHUM
			Scenario 1			
$n = (60, 60, 60)$	1.2	1.045 (-0.155, 0.179)	1.230 (0.030, 0.294)	1.995 (0.795, 0.067)	1.275 (0.075, 0.367)	1.308 (0.108, 0.377)
	1.1	1.018 (-0.082, 0.170)	1.124 (0.024, 0.284)	1.990 (0.890, 0.070)	1.182 (0.082, 0.382)	1.215 (0.115, 0.384)
$n = (90, 90, 90)$	1.2	1.050 (-0.15, 0.125)	1.230 (0.030, 0.229)	1.998 (0.798, 0.062)	1.274 (0.074, 0.297)	1.282 (0.082, 0.311)
	1.1	1.010 (-0.09, 0.113)	1.125 (0.025, 0.219)	1.990 (0.890, 0.062)	1.175 (0.075, 0.289)	1.178 (0.078, 0.299)
$n = (120, 120, 120)$	1.2	1.074 (-0.126, 0.135)	1.219 (0.019, 0.200)	1.994 (0.794, 0.059)	1.256 (0.056, 0.238)	1.258 (0.058, 0.246)
	1.1	1.013 (-0.087, 0.114)	1.117 (0.017, 0.184)	1.973 (0.873, 0.092)	1.144 (0.044, 0.215)	1.148 (0.048, 0.224)
			Scenario 2			
$n = (60, 60, 60)$	1.378	1.059 (-0.320, 0.180)	1.502 (0.124, 0.557)	2.000 (0.622, 0.066)	1.628 (0.25, 0.657)	1.670 (0.292, 0.658)
	1.189	1.006 (-0.183, 0.139)	1.291 (0.102, 0.503)	1.994 (0.805, 0.068)	1.399 (0.21, 0.616)	1.446 (0.257, 0.598)
$n = (90, 90, 90)$	1.378	1.086 (-0.293, 0.218)	1.457 (0.079, 0.447)	2.003 (0.625, 0.087)	1.546 (0.168, 0.549)	1.577 (0.198, 0.526)
	1.189	1.019 (-0.170, 0.174)	1.259 (0.070, 0.395)	1.986 (0.797, 0.081)	1.337 (0.148, 0.484)	1.369 (0.180, 0.479)
$n = (120, 120, 120)$	1.378	1.111 (-0.267, 0.237)	1.414 (0.036, 0.338)	2.006 (0.628, 0.102)	1.474 (0.096, 0.411)	1.485 (0.107, 0.409)
	1.189	1.025 (-0.164, 0.185)	1.216 (0.027, 0.307)	1.977 (0.788, 0.132)	1.272 (0.082, 0.381)	1.282 (0.093, 0.382)
			Scenario 3			
$n = (60, 60, 60)$	1.256	1.058 (-0.199, 0.167)	1.299 (0.042, 0.345)	2.011 (0.754, 0.246)	1.400 (0.144, 0.490)	1.446 (0.189, 0.515)
	0.903	0.964 (0.062, 0.137)	0.947 (0.044, 0.323)	1.983 (1.081, 0.068)	1.032 (0.130, 0.435)	1.086 (0.183, 0.471)
$n = (90, 90, 90)$	1.256	1.102 (-0.154, 0.255)	1.292 (0.036, 0.284)	2.003 (0.746, 0.073)	1.338 (0.082, 0.350)	1.362 (0.106, 0.370)
	0.903	0.957 (0.054, 0.206)	0.932 (0.029, 0.253)	1.977 (1.075, 0.086)	0.974 (0.071, 0.318)	0.993 (0.091, 0.346)
$n = (120, 120, 120)$	1.256	1.122 (-0.135, 0.189)	1.284 (0.028, 0.240)	2.005 (0.748, 0.089)	1.324 (0.067, 0.301)	1.328 (0.071, 0.319)
	0.903	0.940 (0.038, 0.144)	0.917 (0.015, 0.224)	1.965 (1.062, 0.105)	0.948 (0.045, 0.277)	0.951 (0.048, 0.291)
			Scenario 4			
$n = (60, 60, 60)$	0.047	0.695 (0.648, 0.368)	0.964 (0.917, 0.927)	1.960 (1.913, 0.233)	0.089 (0.042, 0.091)	0.100 (0.053, 0.130)
	0.456	1.028 (0.571, 0.538)	3.237 (2.781, 4.707)	3.894 (3.437, 41.735)	0.530 (0.074, 0.225)	0.563 (0.107, 0.276)
$n = (90, 90, 90)$	0.047	0.440 (0.393, 0.387)	1.031 (0.984, 0.817)	1.641 (1.594, 8.895)	0.079 (0.032, 0.055)	0.080 (0.033, 0.061)
	0.456	0.925 (0.469, 0.359)	2.450 (1.993, 1.572)	2.324 (1.868, 8.906)	0.505 (0.049, 0.159)	0.513 (0.056, 0.171)
$n = (120, 120, 120)$	0.047	0.306 (0.259, 0.345)	1.173 (1.126, 1.059)	1.888 (1.841, 0.308)	0.073 (0.025, 0.046)	0.073 (0.026, 0.046)
	0.456	0.838 (0.382, 0.344)	2.548 (2.091, 1.893)	2.020 (1.564, 0.406)	0.492 (0.036, 0.140)	0.494 (0.037, 0.144)

Table 3. Means and standard errors (in parenthesis) of obtained EHUM values at the optimal coefficient vector estimated using the methods: the Empirical method ((10)), the Fréchet bounds method ((12)), the parametric method ((15)), the Min-Max method, SSHUM and NSHUM for simulation Scenarios 1, 2, 3, 4 with sample sizes $(20, 20, 20, 20), (30, 30, 30, 30), (40, 40, 40, 40)$ and $M = 4$, based on 1000 repetitions.

(n_1, n_2, n_3, n_4)	Empirical	Min-Max	Fréchet	SSHUM	NSHUM
		Scenario 1			
(20,20,20,20)	0.846 (0.052)	0.799 (0.059)	0.837 (0.055)	0.871 (0.050)	0.870 (0.051)
(30,30,30,30)	0.848 (0.041)	0.801 (0.047)	0.835 (0.043)	0.871 (0.039)	0.864 (0.040)
(40,40,40,40)	0.847 (0.036)	0.798 (0.041)	0.835 (0.037)	0.859 (0.035)	0.859 (0.035)
		Scenario 2			
(20,20,20,20)	0.712 (0.065)	0.682 (0.067)	0.714 (0.064)	0.746 (0.063)	0.743 (0.064)
(30,30,30,30)	0.714 (0.054)	0.684 (0.055)	0.714 (0.054)	0.737 (0.052)	0.736 (0.053)
(40,40,40,40)	0.715 (0.046)	0.683 (0.047)	0.712 (0.046)	0.732 (0.045)	0.732 (0.045)
		Scenario 3			
(20,20,20,20)	0.765 (0.060)	0.732 (0.065)	0.763 (0.063)	0.798 (0.058)	0.796 (0.058)
(30,30,30,30)	0.766 (0.050)	0.732 (0.053)	0.763 (0.051)	0.789 (0.048)	0.788 (0.048)
(40,40,40,40)	0.767 (0.042)	0.733 (0.044)	0.763 (0.043)	0.783 (0.041)	0.784 (0.041)

(Scenario 4) simulation scenarios. Now both simulation Scenario 1 and 4 can also be generalized for different dimensions of $\mathbf{X}$. To have an idea regarding how the computational time is dependent on the dimension of $\mathbf{X}$ and sample size, we consider Scenario 1 and 4 for sample sizes $(30, 30, 30), (60, 60, 60)$ and dimensions $2, 4, 8$.

Table 4. Means and standard errors (in parenthesis) of obtained EHUM values at the optimal coefficient vector estimated using the methods: the Empirical method ((10)), the Fréchet bounds method ((12)), the parametric method ((15)), the Min-Max method, SSHUM and NSHUM for simulation Scenarios 1, 2, 3 with sample sizes $(30, 30, 30, 30)$ and $M = 4$, based on 1000 repetitions.

(n_1, n_2, n_3, n_4)	Empirical	Min-Max	Fréchet	SSHUM	NSHUM
		Scenario 1			
(20,20,20,20)	0.650 (0.080)	0.411 (0.074)	0.594 (0.078)	0.706 (0.070)	0.697 (0.071)
(30,30,30,30)	0.656 (0.062)	0.415 (0.059)	0.611 (0.066)	0.687 (0.058)	0.682 (0.059)
(40,40,40,40)	0.658 (0.051)	0.411 (0.050)	0.620 (0.058)	0.678 (0.049)	0.674 (0.049)
		Scenario 2			
(20,20,20,20)	0.650 (0.080)	0.411 (0.074)	0.594 (0.078)	0.706 (0.070)	0.697 (0.071)
(30,30,30,30)	0.656 (0.062)	0.415 (0.059)	0.611 (0.066)	0.687 (0.058)	0.682 (0.059)
(40,40,40,40)	0.658 (0.051)	0.411 (0.050)	0.620 (0.058)	0.678 (0.049)	0.674 (0.049)
		Scenario 3			
(20,20,20,20)	0.650 (0.080)	0.411 (0.074)	0.594 (0.078)	0.706 (0.070)	0.697 (0.071)
(30,30,30,30)	0.656 (0.062)	0.415 (0.059)	0.611 (0.066)	0.687 (0.058)	0.682 (0.059)
(40,40,40,40)	0.658 (0.051)	0.411 (0.050)	0.620 (0.058)	0.678 (0.049)	0.674 (0.049)

Here we mainly consider two different optimization methods to maximize the objective functions D_E and D_{s_n}, namely derivative-based Quasi-newton method ((18)) and derivative-free Nelder-Mead (N-M) algorithm ((19)). The Quasi-newton (Q-N) algorithm and the Nelder-Mead algorithms are available under the function options of `fminunc` and `fminsearch` respectively in MATLAB. All these codes have been compiled in MATLAB2016b in a cluster with Intel(R) Xeon(R) CPU E5-2680 v3, 12 Core, 2.5 GHz, 64 bit machines with 256 GB RAM. Each above-mentioned experiments are repeated 100 times and each time different starting points are used. However within each repetition, same starting points are used for all the considered methods. Mean and standard errors of starting point EHUM values, obtained EHUM values (for EVUS and SSHUM) and computation times using both derivative-based and derivative-free methods for simulation scenarios 1 and 4 have been provided in Table 5 and 6 respectively. As mentioned earlier, it is noted that derivative-based Q-N algorithm has completely failed to maximize D_E. For all the considered scenarios, final EHUM values obtained by maximizing D_E using Q-N algorithm are same as the initial point EHUM values. Therefore, we do not consider the maximization of D_E with Q-N algorithm in the following discussion. However, D_{s_n} being smooth, Q-N algorithm has successfully maximized it. In general, no noticeable difference between the obtained EHUM values by maximizing D_{s_n} with Q-N algorithm and N-M algorithm is observed and both of them outperform (in terms of obtained EHUM value) the maximization of D_E with N-M algorithm. But it is noted that maximizing D_{s_n} with Q-N algorithm is in general faster than maximizing D_{s_n} with N-M algorithm which acknowledges the fact that in general derivative-based algorithms are faster than derivative-free algorithms. Upto 2.4 times improvement in computation time is achieved while maximizing D_{s_n} with Q-N algorithm compared to maximizing D_E with N-M algorithm. Specifically, for higher sample size scenarios, maximizing D_{s_n} with Q-N algorithm saves comparatively more computation time than maximizing D_E with N-M algorithm. It should be also

Table 5. Computation times (in seconds), starting values and the final EHUM values at the solution obtained by maximizing D_E (in EVUS) and D_{s_n} (in SSHUM) functions using Quasi-newton (derivative-based) and Nelder-Mead simplex (derivative-free) algorithms for Scenario 1 with sample sizes $(30, 30, 30)$, $(60, 60, 60)$ and $d = 2, 4, 8$.

Dimension	(n_1, n_2, n_3)	Methods	Starting value	D_E(EVUS) (final value)	D_{s_n}(SSHUM) (final value)	D_E(EVUS) (time)	D_{s_n}(SSHUM) (time)
$4*d = 2$	2*(30,30,30)	Quasi-newton (derivative based)	0.635(0.065)	0.635(0.065)	0.672(0.061)	1.86(0.36)	10.11(4.91)
		Nelder-Mead (derivative free)	0.635(0.065)	0.654(0.068)	0.674(0.059)	12.07(4.28)	11.61(3.94)
	2*(60,60,60)	Quasi-newton (derivative based)	0.635(0.050)	0.635(0.050)	0.671(0.043)	123.69(8.74)	749.81(302.32)
		Nelder-Mead (derivative free)	0.635(0.050)	0.668(0.045)	0.672(0.043)	987.51(252.85)	905.37(212.73)
$4*d = 4$	2*(30,30,30)	Quasi-newton (derivative based)	0.802(0.053)	0.802(0.053)	0.850(0.041)	4.10(0.45)	25.69(8.51)
		Nelder-Mead (derivative free)	0.802(0.053)	0.833(0.055)	0.850(0.042)	39.00(9.34)	35.07(10.54)
	2*(60,60,60)	Quasi-newton (derivative based)	0.807(0.042)	0.807(0.042)	0.847(0.030)	406.69(133.25)	2087.83(702.20)
		Nelder-Mead (derivative free)	0.807(0.042)	0.843(0.032)	0.849(0.029)	5008.93(1193.61)	2817.54(677.12)
$4*d = 8$	2*(30,30,30)	Quasi-newton (derivative based)	0.957(0.026)	0.957(0.026)	0.981(0.014)	8.43(0.82)	61.76(16.39)
		Nelder-Mead (derivative free)	0.957(0.026)	0.971(0.020)	0.983(0.012)	82.84(17.83)	79.72(13.14)
	2*(60,60,60)	Quasi-newton (derivative based)	0.958(0.022)	0.958(0.022)	0.979(0.009)	1041.12(153.43)	7496.95(2168.67)
		Nelder-Mead (derivative free)	0.958(0.022)	0.977(0.010)	0.98(0.008)	9630.98(2243.27)	8857.36(2582.59)

Table 6. Computation times (in seconds), starting values and the final EHUM values at the solution obtained by maximizing D_E (in EVUS method) and D_{s_n} (in SSHUM method) functions using Quasi-newton (derivative-based) and Nelder-Mead simplex (derivative-free) algorithms for Scenario 4 with sample sizes $(30, 30, 30)$, $(60, 60, 60)$ and $d = 2, 4, 8$.

Dimension	(n_1, n_2, n_3)	Methods	Starting value	D_E(EVUS) (final value)	D_{s_n}(SSHUM) (final value)	D_E(EVUS) (time)	D_{s_n}(SSHUM) (time)
$4*d = 2$	2*(30,30,30)	Quasi-newton (derivative based)	0.327(0.069)	0.327(0.069)	0.354(0.069)	1.72(0.18)	9.04(5.03)
		Nelder-Mead (derivative free)	0.327(0.069)	0.347(0.073)	0.358(0.067)	11.84(3.60)	10.22(3.16)
	2*(60,60,60)	Quasi-newton (derivative based)	0.330(0.053)	0.330(0.053)	0.352(0.050)	131.47(9.56)	775.72(363.94)
		Nelder-Mead (derivative free)	0.330(0.053)	0.357(0.051)	0.361(0.048)	941.9(257.68)	838.05(254.42)
$4*d = 4$	2*(30,30,30)	Quasi-newton (derivative based)	0.551(0.123)	0.551(0.123)	0.908(0.036)	3.86(0.34)	35.77(10.08)
		Nelder-Mead (derivative free)	0.551(0.123)	0.888(0.093)	0.911(0.033)	43.14(9.48)	38.01(6.40)
	2*(60,60,60)	Quasi-newton (derivative based)	0.544(0.110)	0.544(0.110)	0.901(0.025)	269.97(18.86)	2679.7(661.86)
		Nelder-Mead (derivative free)	0.544(0.110)	0.901(0.026)	0.902(0.024)	2949.92(419.30)	3118.1(455.82)
$4*d = 8$	2*(30,30,30)	Quasi-newton (derivative based)	0.773(0.124)	0.773(0.124)	1.000(0.004)	8.57(0.97)	33.56(13.08)
		Nelder-Mead (derivative free)	0.773(0.124)	0.994(0.016)	1.000(0.003)	87.22(21.43)	99.43(19.38)
	2*(60,60,60)	Quasi-newton (derivative based)	0.774(0.113)	0.774(0.113)	1.000(0.002)	564.66(39.54)	3240.6(1210.61)
		Nelder-Mead (derivative free)	0.774(0.113)	0.999(0.001)	1.000(0.000)	7194.59(1416.38)	6860.81(1034.17)

noted that the computation time required for maximizing D_{s_n} with Q-N algorithm method varies less compared to the computation time required for maximizing D_E with N-M algorithm.

6. Data Analysis

The Alzheimer's Disease Data Analysis

For illustration of our methodology, we analyze a subset of the longitudinal cohort data on Alzheimer's Disease (AD) from Alzheimer's Disease Research Center (ADRC) at the University of Washington. The dataset is available in the R package `DiagTest3Grp`. In this data set, measurements of 14 neuro-psychological markers were collected from 118 independent individuals of age 75 and above among which 44 individuals were labeled as non-demented, 43 were mildly demented, and 21 individuals were labeled as demented, i.e., Alzheimer's disease. It is now commonly accepted that treatment for Alzheimer's disease is a rather complicated process and a more clinically useful strategy is to apply appropriate interventions for earlier stage patients with relatively mild conditions (see (20), (21)). Therefore it is meaningful to differentiate three or even more categories of patients with ascending disease severity and subsequently offer category-specific treatments.

Due to some missing observations, we delete 10 individuals from the data set for our analysis. Note that values of these fourteen biomarkers can be negative. Furthermore, as we can see from the boxplot in Figure 3 and density plot in Figure 4, there is a clear decreasing trend in the distributions of the neuro-psychological markers across the dementia status, except for *zbentd*. This shows the potential discrimination power of the other thirteen individual markers. This observation is further reinforced by their individual discrimination power in terms of EHUM values where *factor1*, *ktemp* and *zpsy004* have the highest individual EHUM values ranging from 0.70 to 0.78. However, the EHUM value for *zbentd* is 0.144, slightly smaller than the lowest EHUM value for random guess which is 0.17 in this case. These values are presented in Table 7. Recall that for random guess the HUM value is 1/6=0.1667 when the disease outcome variable has three possible categories. That is to say that a HUM value for any biomarker less than 0.1667 indicates that the biomarker is weaker than random guessing in predicting the disease outcome and should be avoided from the prediction model. As a result, we excluded *zbentd* from the subsequent analysis.

To see the improvement in discrimination accuracy by combining these individual markers over the individual markers and to facilitate comparison, we employ all the six combining methods discussed in Section 3. For all the six methods, the estimated EHUM values with their respective standard errors are reported in Table 8. We also reported the coefficient parameter estimates with their respective bootstrap standard errors. As we can see, SSHUM gives the highest EHUM of 0.874, followed by NSHUM (0.849), empirical (0.832). However, the Min-Max and the Naive method have the lowest EHUM values of 0.80 and 0.792, respectively.

In practice, the concerns about overfitting motivate us to select a subset of biomarkers for subsequent combination. Cross-validation offers a simple way for tuning the number of biomarkers. Here, we employed a 5-fold cross-validation to compare the discrimination accuracy of varying number of biomarkers after combination. For simplicity, we only implemented the SSHUM and NSHUM

approaches. In the analysis, the biomarkers were ranked according to their individual HUMs. For example, if the number of biomarkers were 2, then *ktemp* and *FACTOR1* would be selected. Table 9 presents the EHUM values with different subsets of the 13 biomarkers. For the SSHUM method, combining the first 4 biomarkers (i.e., *ktemp*, *FACTOR1*, *zpsy004*, and *kfront*) gives the highest EHUM value, while for the NSHUM method, combining the first 5 biomarkers (i.e., *ktemp*, *FACTOR1*, *zpsy004*, *kfront*, and *zassc*) gives the highest EHUM value. These findings may indicates the importance of selecting biomarkers before combination, which could be a direction for future work.

The National Institute of Aging-Alzheimer's Association (NIA-AA) published research criteria for AD diagnosis in 2011 using biomarkers information. In addition to dementia due to AD, other stages of interest include prodromal AD (mild cognitive impairment) and preclinical AD (individuals with normal condition with AD pathology). The markers evaluated in our analysis may also offer useful insight for such mutli-stage diagnosis.

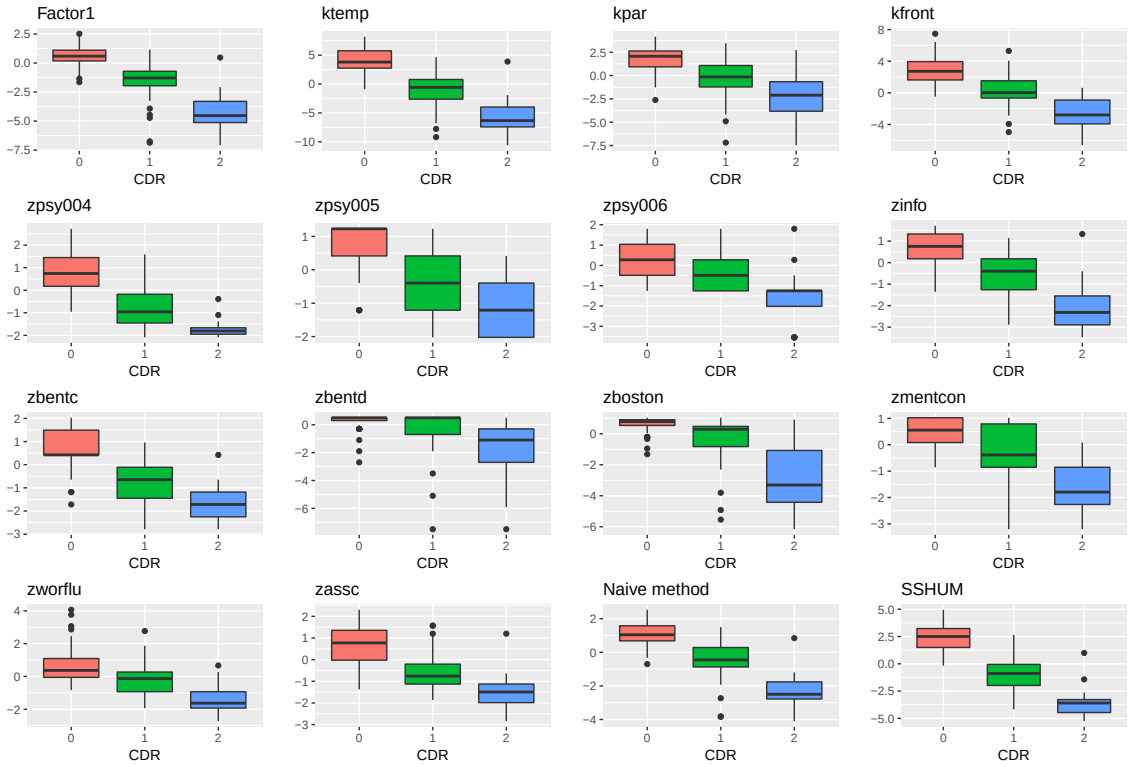

Figure 3. Boxplot for individual and combined biomarkers for Alzheimer data set.

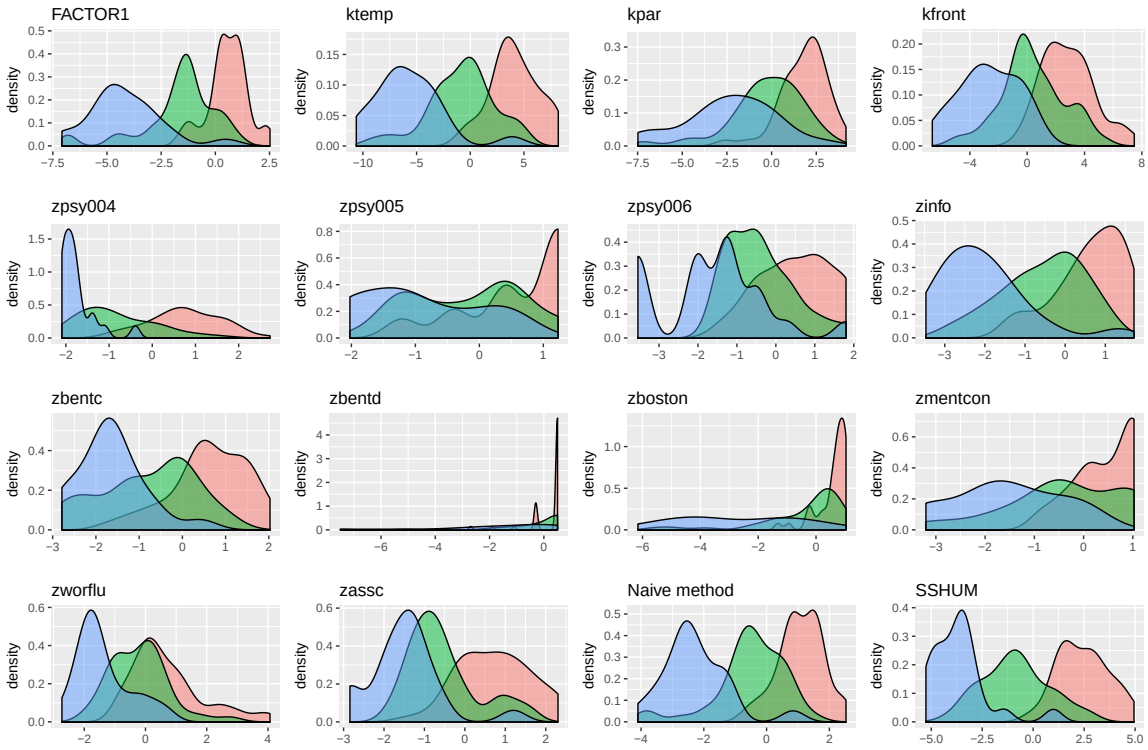

Figure 4. Density plot for individual and combined biomarkers for Alzheimer data set.

Table 7. Empirical HUM values (with bootstrap standard errors) for the individual biomarkers for the AKI and Alzheimer data sets.

Alzheimer data		ERICCA data	
Individual biomarkers	HUM (se)	Individual biomarkers	HUM (se)
FACTOR1	0.774 (0.056)	NGAL 0 hours	0.179 (0.029)
ktemp	0.784 (0.055)	NGAL 6 hours	0.222 (0.034)
kpar	0.600 (0.065)	NGAL 12 hours	0.273 (0.040)
kfront	0.654 (0.059)	NGAL 24 hours	0.315 (0.042)
zpsy004	0.718 (0.058)		
zpsy005	0.316 (0.064)		
zpsy006	0.442 (0.069)		
zinfo	0.643 (0.065)		
zbentc	0.506 (0.060)		
zbentd	0.144 (0.047)		
zboston	0.590 (0.066)		
zmentcon	0.367 (0.065)		
zworflu	0.561 (0.066)		
zassc	0.648 (0.066)		

The ERICCA Trial Data Analysis

Here we analyze an acute kidney injury (AKI) dataset following a heart surgery to illustrate our proposed method. We consider the data from the **E**ffect of

Table 8. Estimated coefficients and the HUM values (with standard errors in parenthesis) for the Alzheimer's disease data using Naive method, Empirical, SSHUM, NSHUM, Fréchet, Parametric and Min-max methods.

Biomarkers	β_{Naive}	$\beta_{Empirical}$	β_{SSHUM}	β_{NSHUM}	$\beta_{Frechet}$	$\beta_{Parametric}$	$\beta_{Min-Max}$
FACTOR1	0.077	0.271 (0.004)	-0.229 (0.248)	0.300 (0.238)	0.272 (0.002)	0.277 (0.041)	-
ktemp	0.077	0.283 (0.012)	0.335 (0.144)	0.364 (0.133)	0.281 (0.011)	0.277 (0.075)	-
kpar	0.077	0.251 (0.018)	0.013 (0.149)	0.090 (0.168)	0.249 (0.009)	0.277 (0.044)	-
kfront	0.077	0.283 (0.017)	-0.026 (0.143)	0.107 (0.140)	0.281 (0.012)	0.277 (0.062)	-
zpsy004	0.077	0.283 (0.007)	0.469 (0.134)	0.461 (0.122)	0.281 (0.002)	0.277 (0.073)	-
zpsy005	0.077	0.275 (0.006)	0.160 (0.145)	0.236 (0.152)	0.281 (0.003)	0.277 (0.053)	-
zpsy006	0.077	0.283 (0.013)	0.469 (0.142)	0.465 (0.132)	0.281 (0.009)	0.277 (0.049)	-
zinfo	0.077	0.283 (0.006)	-0.454 (0.224)	0.008 (0.221)	0.281 (0.003)	0.277 (0.057)	-
zbentc	0.077	0.269 (0.013)	0.312 (0.190)	0.401 (0.175)	0.271 (0.009)	0.277 (0.008)	-
zboston	0.077	0.283 (0.005)	0.057 (0.223)	0.159 (0.195)	0.281 (0.003)	0.277 (0.027)	-
zmentcon	0.077	0.283 (0.009)	0.194 (0.197)	0.290 (0.174)	0.281 (0.005)	0.277 (0.014)	-
zworflu	0.077	0.283 (0.008)	0.091 (0.189)	0.058 (0.160)	0.281 (0.002)	0.277 (0.011)	-
zassc	0.077	0.274 (0.007)	-0.126 (0.221)	-0.008 (0.223)	0.281 (0.003)	0.277 (0.009)	-
HUM	0.799	0.801 (0.059)	0.851 (0.058)	0.823 (0.058)	0.801 (0.058)	0.799 (0.097)	0.804 (0.058)

Table 9. Empirical HUM values for the Alzheimer's disease data with varying number of combined biomarkers using the SSHUM and NSHUM methods.

SSHUM		NSHUM	
Number of biomarkers	HUM	Number of biomarkers	HUM
1	0.879	1	0.879
2	0.912	2	0.898
3	0.913	3	0.913
4	**0.927**	4	0.899
5	**0.927**	5	**0.918**
6	0.885	6	0.877
7	0.810	7	0.860
8	0.810	8	0.835
9	0.799	9	0.851
10	0.860	10	0.843
11	0.887	11	0.903
12	0.748	12	0.767
13	0.798	13	0.810

Remote **I**schemic Preconditioning on **C**linical Outcomes in Patient Undergoing **C**oronary **A**rtery Bypass Graft Surgery (ERICCA) trial where a group of 1612 patients participated in a cardiovascular surgery and were observed for one year after the surgery ((22; 16)). All the patients were randomized to two different methods of surgeries namely Remote Ischemic Conditioning (RIC) or Sham Preconditioning. During the study period, some patients developed AKI along with few other diseases post-surgery. The AKI was recorded as a multi-category ordinal outcome with four levels based on the severity level. The data also includes cardiovascular death and all-cause mortality at 1 year (binary), non-fatal Myocardial Infarction (MI) (binary) and coronary revascularization or stroke at 1 year (binary). In the literature, studies on prediction of AKI after cardiac surgery has been performed in several occasions. Assuming AKI as a binary outcome, (23) found that the serum Neutrophil Gelatinase Associated Lipocalin (NGAL) measurements taken at 0 (before surgery), 6, 12 and 24 hours after surgery are significant influential biomarkers in the development of AKI. In addition, they showed that for the risk-stratification of patients prior to cardiac surgery for AKI may be improved by adding pre-oprative levels of NGAL to existing risk scores where existing risk score was calculated based on age, gender,

diabetes mellitus, hypertension, peripheral vascular disease, previous Coronary Artery Bypass Graft type of surgery planned, use of intra-aortic ballon pump and few other baseline covariates. However, the main limitation of their study is that they did not consider the multiple categories of the AKI outcome. Instead, they converted it to a binary outcome where level 0 stands for no AKI and level 1 stands for any of the 1, 2, 3 levels of AKI in the data.

To illustrate the proposed method, we consider the AKI within 72 hours of surgery as our multi-category outcome which are leveled as 0 (none), 1, 2, 3 as per the international Kidney Disease: Improving Global Outcomes classification (KDIGO) criteria on serum creatinine. Since level 3 has only a few observations, we combine the levels 2 and 3 into a single category denoted as the highest risk group. Therefore, in the following analysis, the AKI has three categories. Our biomarkers of interest in predicting AKI are individual NGAL at 0 (before surgery), 6, 12 and 24 hours after surgery and their different combinations using different methods. In a previous analysis, (23) observed that there is a significant increase in AKI as the individual's pre-operative NGAL increases from the first to the third tertile (>220 ng/L). Hence they considered only the individuals from the third tertile and concluded that the pre-operative NGAL is a significant predictor in predicting binary AKI. There are 305 individuals in our sample after discarding all the missing observations. Among these subjects, 172 patients did not develop AKI within the 72 hours of surgery (AKI=0), 99 patients developed level 1 AKI, and 34 developed level 2 (i.e., combined levels 2 and 3 in original scale) AKI.

Note that larger values of the NGAL measurements indicate higher levels of severity of AKI. Since the NGAL measurements are highly skewed-distributed and large in number, we transform them into the logarithm scale to scale down those high numbers and make the distributions close to normal distributions. Considering logarithmic transformation of the biomarkers is a common strategy for this type of data analysis (see e.g., (2)). To see the visual discrimination power of these individual log of NGAL measurements, the box plots and the density plots are shown in Figures 5 and 6, respectively. The estimated empirical HUM values for the individual NGAL at four different time points are 0.179 (at 0 hours), 0.222 (at 6 hours), 0.273 (at 12 hours), and 0.315 (at 24 hours), respectively, clearly much larger than the lowest EHUM value for random guess which is 0.17 in this case. Recall that for random guess the HUM value is 1/6=0.1667 when the disease outcome variable has three possible categories. These values are also reported in Table 7, along with their respective standard errors. In this case, all the NGAL measurements can be included in the prediction model. Further, it is worth noting that as the time of NGAL measurement increases from 0 hours to 24 hours, the HUM value increases to almost two times that of the 0 hours. It indicates the strong discrimination power of the NGAL biomarker in predicting AKI as time progresses after surgery.

Further, we treat the four NGAL measurements as four biomarkers and apply our proposed SSHUM method to combine these markers. As comparison, a naive linear combination approach with equal weights on the four markers is also constructed. The distributions of these combined markers are also displayed in

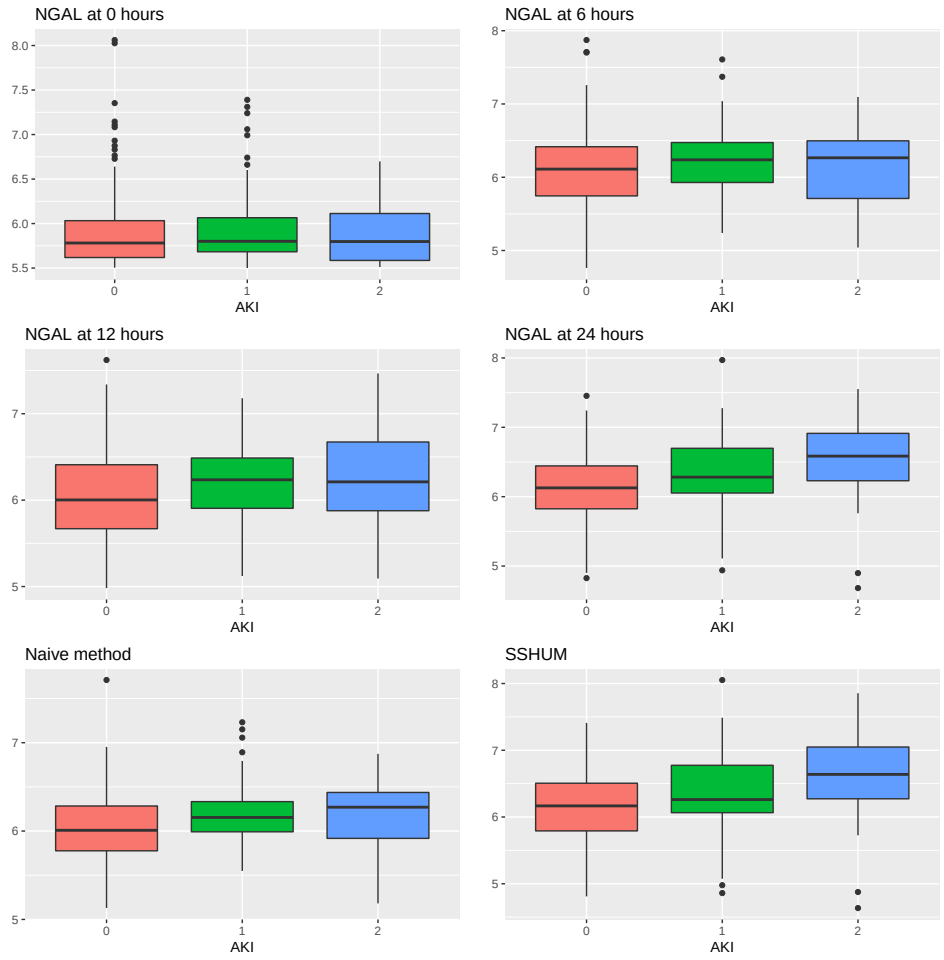

Figure 5. Boxplot for individual and combined NGALs for ERICCA data set. The top 4 plots represents the NGAL levels at 0, 6, 12 and 24 hours after the surgery for 3 levels of AKI. Bottom left diagram shows the boxplots for Naive method (i.e., linear combination of covariates with equal positive coefficients) and the bottom right diagram shows the boxplots for SSHUM method.

Figures 5 and 6. It is noted that SSHUM separates the three class in the most effective way.

Further, we obtain the HUM values for other existing methods along with their respective optimal linear combination estimates. The estimates along with their bootstrap standard errors are reported in Table 10. We note that all the linear combining methods yield larger HUM values than that of the individual biomarkers and the naive equal weight method. The proposed sigmoid approximation yields the highest HUM value compared to the other existing methods. Although the proposed method combines the time-varying NGAL

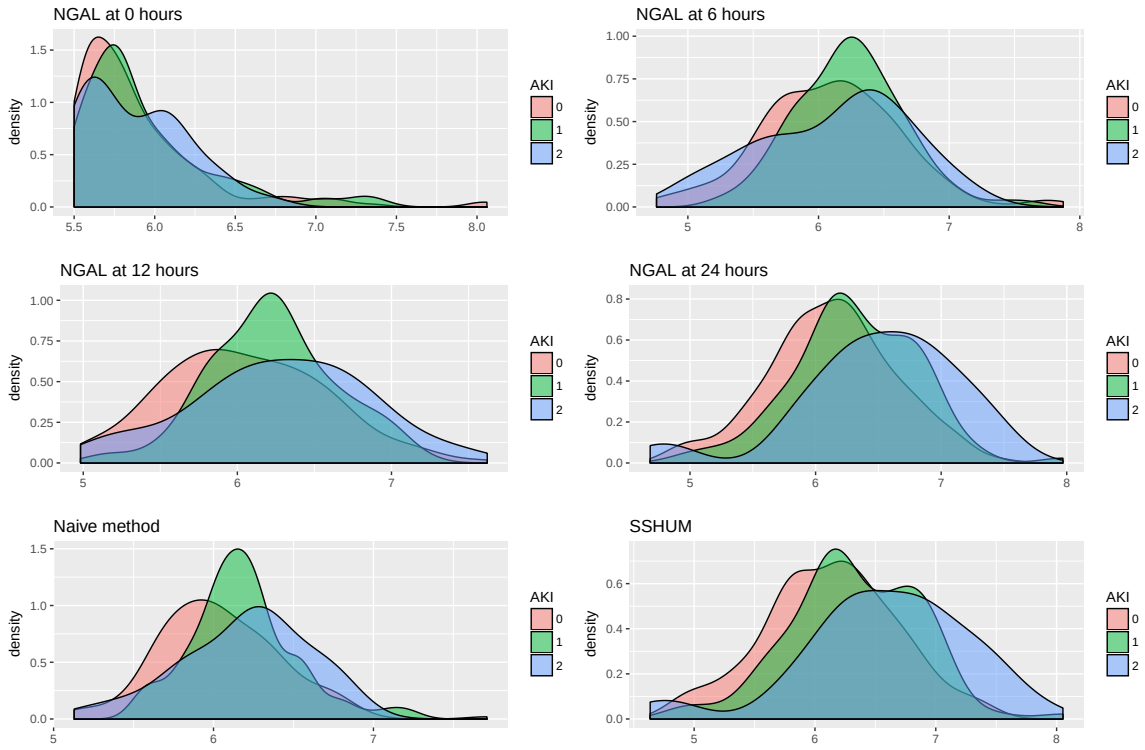

Figure 6. Density plot for individual and combined NGALs for ERICCA data set.

measurements in a more effective way than the others, further research may be required to support the effectiveness of such NGAL measurements and their combining factor in predicting AKI.

Table 10. Estimated optimal coefficients and the HUM values (with standard errors in parenthesis) for the ERICCA dataset using naive method, empirical method, SSHUM, NSHUM, Fréchet, parametric and Min-Max methods.

Biomarkers	*naive*	*Empirical*	*SSHUM*	*NSHUM*	*Frechet*	*Parametric*	$Min-Max$
NGAL 0 hours	0.5	0.412 (0.2869)	-0.208 (0.3078)	-0.097 (0.3142)	0.234 (0.0798)	0.236 (0.3636)	-
NGAL 6 hours	0.5	-0.050 (0.4074)	-0.660 (0.3201)	-0.387 (0.3196)	-0.382 (0.1083)	0.593 (0.3659)	-
NGAL 12 hours	0.5	0.594 (0.2098)	0.360 (0.3320)	0.176 (0.3377)	0.566 (0.0426)	0.590 (0.2563)	-
NGAL 24 hours	0.5	0.688 (0.1917)	1.508 (0.2570)	0.900 (0.2665)	0.692 (0.0462)	0.494 (0.1762)	-
HUM	0.281	0.317 (0.0154)	0.326 (0.0140)	0.325 (0.0135)	0.312 (0.0054)	0.287 (0.0182)	0.303 (0.0079)

7. Discussion

Improving diagnostic accuracy by combining multiple biomarkers has been studied for binary and multi-category outcomes. In this article, we have extended the idea of direct maximization of the empirical HUM, specifically the VUS proposed by (10), to a smoothing approximation using a class of smooth CDFs

which can be further controlled by a tuning parameter λ. In particular, we have considered the logistic CDF (sigmoid function) and normal CDF (probit function) to operationalize our proposed method. We have also discussed about the choice of the tuning parameter λ using bias-variance trade-off. Consistency and asymptotic normality of the coefficient estimators using the proposed method have been established. Through simulation studies we have observed that the proposed method is computationally less challenging than the direct maximization of the empirical HUM which is non-smooth and non-differentiable. We also noted that the performance of the proposed method heavily depends on the choice of the tuning parameter λ, with lower values of λ leading to results very similar to the empirical method with less bias but large variability. This is a problem of bias-variance trade-off which we have discussed in considerable detail in Section 2.3. Results from our simulation study have shown that the proposed method outperforms the other existing approximation methods including the empirical HUM method because of the computational issue. We see several future directions for this article. First, within this work we have employed the step-down algorithm to maximize our objective functions. A limitation with this algorithm is that it maximizes the objective function with respect to a single parameter at a time which might result in local optimal solution rather than the global optimal solution. However, in future, coming up with advanced computational tools and fast global optimization algorithms for simultaneous estimation of the whole coefficient vector (instead of estimating one at a time using step-down algorithm) maximizing the objective function might further improve the solutions. In addition, there might be ties in the markers under study. Considering the possibility of ties can potentially reduce the number of coefficient parameters and lead to a more efficient estimator.

Acknowledgements

We thank the three reviewers and the associate editor for their help comments which led to this improved version of the paper. We also thank Jon Wellner, Palash Ghosh and Heerajnarain Bulluck for helpful discussions. The work was partially supported by grants R-155-000-205-114, R-155-000-195-114, R-155-000-197-112, R-155-000-197-113 and MOE2015-T2-2-056 from the Ministry of Education in Singapore, as well as the start-up grant of Bibhas Chakraborty from Duke-NUS Medical School. The ERICCA trial was funded by the Efficacy and Mechanism Evaluation Program, a Medical Research Council and National Institute of Health Research partnership, and the British Heart Foundation [Grant Reference number 09/100/05]. We would like to thank the patients and research team of the ERICCA trial, and The London School of Hygiene and Tropical Medicine Clinical Trials Unit, UK who coordinated the trial. Derek Hausenloy was supported by the British Heart Foundation (CS/14/3/31002), the Duke-NUS Signature Research Programme funded by the Ministry of Health, Singapore Ministry of Health's National Medical Research Council under its Clinician Scientist-Senior Investigator scheme (NMRC/CSA-SI/0011/2017), Centre Grant, and Collaborative Centre Grant scheme (NMRC/CGAug16C006). This article is based upon work from COST

Action EU-CARDIOPROTECTION CA16225 supported by COST (European Cooperation in Science and Technology).

References

[1] Su JQ and Liu JS. Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association* 1993; 88(424): 1350–1355.

[2] Pepe MS and Thompson ML. Combining diagnostic test results to increase accuracy. *Biostatistics* 2000; 1(2): 123–140.

[3] Pepe MS, Cai T and Longton G. Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics* 2006; 62(1): 221–229.

[4] Ma S and Huang J. Combining multiple markers for classification using ROC. *Biometrics* 2007; 63(3): 751–757.

[5] Liu C, Liu A and Halabi S. A min–max combination of biomarkers to improve diagnostic accuracy. *Statistics in Medicine* 2011; 30(16): 2005–2014.

[6] Li J and Fine JP. ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies. *Biostatistics* 2008; 9(3): 566–576.

[7] Scurfield BK. Multiple-event forced-choice tasks in the theory of signal detectability. *Journal of Mathematical Psychology* 1996; 40(3): 253–269.

[8] Mossman D. Three-way ROCs. *Medical Decision Making* 1999; 19(1): 78–89.

[9] Nakas CT and Yiannoutsos CT. Ordered multiple-class ROC analysis with continuous measurements. *Statistics in Medicine* 2004; 23(22): 3437–3449.

[10] Zhang Y and Li J. Combining multiple markers for multi-category classification: An ROC surface approach. *Australian & New Zealand Journal of Statistics* 2011; 53(1): 63–78.

[11] Kang L, Xiong C, Crane P et al. Linear combinations of biomarkers to improve diagnostic accuracy with three ordinal diagnostic categories. *Statistics in Medicine* 2013; 32(4): 631–643.

[12] Hsu MJ and Chen YH. Optimal linear combination of biomarkers for multi-category diagnosis. *Statistics in Medicine* 2016; 35(2): 202–213.

[13] Novoselova N, Beffa CD, Wang J et al. HUM calculator and HUM package for R: easy-to-use software tools for multicategory receiver operating characteristic analysis. *Bioinformatics* 2013; 30(11): 1635–6.

[14] Li J, Gao M and D'Agostino R. Evaluating classification accuracy for modern learning approaches. *Statistics in Medicine* 2019; : In press.

[15] Zhang Y. ROC analysis in diagnostic medicine (Phd Thesis). Department of Statistics and Applied Probability, National University of Singapore 2010; .

[16] Hausenloy DJ, Candilio L, Evans R et al. Remote ischemic preconditioning and outcomes of cardiac surgery. *New England Journal of Medicine* 2015; 373(15): 1408–1417.

[17] Gammerman A. *Computational learning and probabilistic reasoning*. John Wiley & Sons, Inc., 1996.

[18] Fletcher R. *Practical methods of optimization*. John Wiley & Sons, 1987.

[19] Nelder J and Mead R. A simplex method for function minimization. *Computer Journal* 1965; 7: 308–313.

[20] Dubois B, Feldman H, Jacova C et al. Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. *Lancet Neurol* 2014; 13(6): 614–29.

[21] Dubois B, Hampel H, Feldman H et al. Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria. *Alzheimers Dement* 2016; 12(3): 292–323.

[22] Hausenloy DJ, Candilio L, Laing C et al. Effect of remote ischemic preconditioning on clinical outcomes in patients undergoing coronary artery bypass graft surgery (ERICCA): rationale and study design of a multicentre randomized double-blinded controlled clinical trial. *Clinical Research in Cardiology* 2012; 101(5): 339–348.

[23] Bulluck H, Maiti R, Chakraborty B et al. Neutrophil gelatinase-associated lipocalin prior to cardiac surgery predicts acute kidney injury and mortality. *Heart* 2018; 104(4): 313–317.

[24] Han AK. Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics* 1987; 35(2-3): 303–316.

[25] Kowalski J and Tu X. *Modern Applied U-Statistics*. John Wiley & Sons, Inc., 2008.

Appendix

A1: Proof of Theorem 1

Assuming (A1)-(A3), (10) proved the consistency of $\widehat{\boldsymbol{\beta}}_E$, an empirical HUM based estimator of $\boldsymbol{\beta}$ for three-category ordinal outcome, using the result of

maximum rank correlation type estimators by (24) . In fact, it can be shown that $\widehat{\boldsymbol{\beta}}_E$ is a consistent estimator of $\boldsymbol{\beta}$ for any number of categories. The above result is equivalent to

$$\sup_{\boldsymbol{\beta}\in B} D_E(\boldsymbol{\beta}) - D(\boldsymbol{\beta}) = o_p(1),$$

i.e., $\sup_{\boldsymbol{\beta}\in B} D_E(\boldsymbol{\beta}) - D(\boldsymbol{\beta})$ converges to 0 in probability.

Similarly, to prove the probability convergence of $\widehat{\boldsymbol{\beta}}_{s_n}$, the proposed SSHUM based estimator, we have to show that

$$\sup_{\boldsymbol{\beta}\in B} D_{s_n}(\boldsymbol{\beta}) - D(\boldsymbol{\beta}) = o_p(1).$$

Note that, using the triangular inequality, we can write

$$\begin{aligned}
\sup_{\boldsymbol{\beta}\in B} D_{s_n}(\boldsymbol{\beta}) - D(\boldsymbol{\beta}) &= \sup_{\boldsymbol{\beta}\in B} D_{s_n}(\boldsymbol{\beta}) - D_E(\boldsymbol{\beta}) + D_E(\boldsymbol{\beta}) - D(\boldsymbol{\beta}), \\
&\leq \sup_{\boldsymbol{\beta}\in B} D_{s_n}(\boldsymbol{\beta}) - D_E(\boldsymbol{\beta}) + \sup_{\boldsymbol{\beta}\in B} D_E(\boldsymbol{\beta}) - D(\boldsymbol{\beta}) \\
&= \sup_{\boldsymbol{\beta}\in B} D_{s_n}(\boldsymbol{\beta}) - D_E(\boldsymbol{\beta}) + o_p(1). \qquad (7)
\end{aligned}$$

Hence, to prove the consistency of $\widehat{\boldsymbol{\beta}}_{s_n}$, it is sufficient to prove the following lemma.

Lemma 1. *Under the assumptions (A1)-(A3),*

$$\sup_{\boldsymbol{\beta}\in B} D_E(\boldsymbol{\beta}) - D_{s_n}(\boldsymbol{\beta}) \xrightarrow{p} 0$$

as $n \to \infty$.

Proof of Lemma 1

For binary outcome, (4) proved the consistency of $\boldsymbol{\beta}_{s_n}$ by showing that

$$\sup_{\boldsymbol{\beta}\in B} D_{s_n}(\boldsymbol{\beta}) - D_E(\boldsymbol{\beta}) = o_p(1).$$

Here, we use the same idea to prove that $\sup_{\boldsymbol{\beta}\in B} D_{s_n}(\boldsymbol{\beta}) - D_E(\boldsymbol{\beta}) = o_p(1)$ for multi-category ordinal outcome. Define an equivalent definition of $D_E(\boldsymbol{\beta})$ as

$$D_E(\boldsymbol{\beta}) = C \sum_{i_1\neq i_2\neq\cdots\neq i_M} I(Y_{Mi_M} > \cdots > Y_{1i_1}) I(\boldsymbol{\beta}^T \mathbf{Z}_{i_M i_{M-1}} > 0) I(\boldsymbol{\beta}^T \mathbf{Z}_{i_{M-1} i_{M-2}} > 0) \cdots I(\boldsymbol{\beta}^T \mathbf{Z}_{i_2 i_1} > 0),$$

where $C = \dfrac{1}{n(n-1)\cdots(n-M+1)}$, $\mathbf{Z}_{i_{j+1}i_j} = \mathbf{X}_{(j+1)i_{j+1}} - \mathbf{X}_{ji_j}$, and Y_{ji_j}, $j = 1, 2, \cdots, M$ are defined as $Y_{ji_j} = j$ if the i_j-th observation belongs to the j-th category, otherwise 0.

Similarly, we define an equivalent definition of SSHUM as

$$D_{s_n}(\boldsymbol{\beta}) = C \sum_{i_1 \neq i_2 \neq \cdots \neq i_M} I(Y_{Mi_M} > \cdots > Y_{1i_1}) s_n(\boldsymbol{\beta}^T \mathbf{Z}_{i_M i_{M-1}}) s_n(\boldsymbol{\beta}^T \mathbf{Z}_{i_{M-1} i_{M-2}}) \cdots s_n(\boldsymbol{\beta}^T \mathbf{Z}_{i_2, i_1})$$

For any $\delta > 0$, we can write

$$D_E(\boldsymbol{\beta}) - D_{s_n}(\boldsymbol{\beta}) \leq T_{n1} + T_{n2}$$

where

$$\begin{aligned} T_{n1} = & C \sum_{i_1 \neq i_2 \neq \cdots \neq i_M} I(Y_{Mi_M} > \cdots > Y_{1i_1}) \\ & I(\boldsymbol{\beta}^T \mathbf{Z}_{i_M i_{M-1}} > 0) \cdots I(\boldsymbol{\beta}^T \mathbf{Z}_{i_2 i_1} > 0) - s_n(\boldsymbol{\beta}^T \mathbf{Z}_{i_M i_{M-1}}) \cdots s_n(\boldsymbol{\beta}^T \mathbf{Z}_{i_2 i_1}) \\ & I\left(\max_{1 \leq j \leq M-1} \boldsymbol{\beta}^T \mathbf{Z}_{i_{j+1} i_j} \geq \delta \right) \end{aligned}$$

and

$$\begin{aligned} T_{n2} = & C \sum_{i_1 \neq i_2 \neq \cdots \neq i_M} I(Y_{Mi_M} > \cdots > Y_{1i_1}) \\ & I(\boldsymbol{\beta}^T \mathbf{Z}_{i_M i_{M-1}} > 0) \cdots I(\boldsymbol{\beta}^T \mathbf{Z}_{i_2 i_1} > 0) - s_n(\boldsymbol{\beta}^T \mathbf{Z}_{i_M i_{M-1}}) \cdots s_n(\boldsymbol{\beta}^T \mathbf{Z}_{i_2 i_1}) \\ & I\left(\max_{1 \leq j \leq M-1} \boldsymbol{\beta}^T \mathbf{Z}_{i_{j+1} i_j} < \delta \right). \end{aligned}$$

(4) showed that on the set $\{x \geq \delta\}$, $s_n(x) - I(x > 0) \leq \exp(-x/\sigma_n) < \exp(-\delta/\sigma_n) \to 0$ uniformly as $\sigma_n \to 0$. Following this, it can be shown that

$$s_n(x_1) \to I(x_1 > 0) \text{ uniformly on the set } \{x_1 \geq \delta\},$$

$$s_n(x_2) \to I(x_2 > 0) \text{ uniformly on the set } \{x_2 \geq \delta\},$$

$$\vdots$$

$$\text{and } s_n(x_{M-1}) \to I(x_{M-1} > 0) \text{ uniformly on the set } \{x_{M-1} \geq \delta\}.$$

It implies that on the set $\{\max_{1\leq i\leq M-1} x_i \geq \delta\}$, $s_n(x_i) \to I(x_i > 0)$ uniformly for all $i = 1, 2, \cdots, M-1$. Following this, we can write

$$\begin{aligned}
& s_n(x_1)s_n(x_2)\cdots s_n(x_{M-1}) - I(x_1>0)I(x_2>0)\cdots I(x_{M-1}>0) \\
\leq\ & s_n(x_1) - I(x_1>0)s_n(x_2)\cdots s_n(x_{M-1}) + \\
& I(x_1>0)s_n(x_2)\cdots s_n(x_{M-1}) - I(x_2>0)\cdots I(x_{M-1}>0), \\
\leq\ & s_n(x_1) - I(x_1>0)s_n(x_2)\cdots s_n(x_{M-1}) + \\
& I(x_1>0)s_n(x_2) - I(x_2>0)s_n(x_3)\cdots s_n(x_{M-1}) + \\
& I(x_1>0)I(x_2>0)s_n(x_3)\cdots s_n(x_{M-1}) - I(x_3>0)\cdots I(x_{M-1}>0), \\
& \vdots \\
\leq\ & s_n(x_1) - I(x_1>0)s_n(x_2)\cdots s_n(x_{M-1}) + \\
& I(x_1>0)s_n(x_2) - I(x_2>0)s_n(x_3)\cdots s_n(x_{M-1}) + \cdots + \\
& I(x_1>0)I(x_2>0)\cdots I(x_{M-2}>0)s_n(x_{M-1}) - I(x_{M-1}>0), \\
=\ & o_p(1) + o_p(1) + \cdots + o_p(1) = o_p(1).
\end{aligned}$$

Now replacing x_j by $\boldsymbol{\beta}^T\mathbf{Z}_{i_{j+1},i_j}$ in the above derivation, we can see that T_{n1} converges to 0 uniformly on set B. The second term can be bounded above as

$$T_{n2} \leq C \sum_{i_1\neq i_2\neq\cdots\neq i_M} I\left(\max_{1\leq j\leq M-1} \boldsymbol{\beta}^T\mathbf{Z}_{i_{j+1}i_j} < \delta\right).$$

Again by the uniform convergence of the U-process, the right hand side of the above equation converges to $P\left(\max_{1\leq j\leq M-1}\boldsymbol{\beta}^T\mathbf{Z}_{i_{j+1}i_j} < \delta\right)$ almost surely on B. Further, using order statistic result, we can write

$$\begin{aligned}
P\left(\max_{1\leq j\leq M-1} \boldsymbol{\beta}^T\mathbf{Z}_{i_{j+1}i_j} < \delta\right) &= P\left(\boldsymbol{\beta}^T\mathbf{Z}_{i_Mi_{M-1}} < \delta, \boldsymbol{\beta}^T\mathbf{Z}_{i_{M-1}i_{M-2}} < \delta, \cdots, \boldsymbol{\beta}^T\mathbf{Z}_{i_2i_1} < \delta\right) \\
&\leq P\left(\boldsymbol{\beta}^T\mathbf{Z}_{i_{j+1}i_j} < \delta\right)
\end{aligned}$$

for all $j = 1, 2, \cdots, M-1$ over B. Under the assumptions (A2) and (A3), it can be shown that $P\left(\boldsymbol{\beta}^T\mathbf{Z}_{i_{j+1}i_j} < \delta\right)$ converges to 0 uniformly over B as δ goes to 0. Hence, it proves that $\sup_{\boldsymbol{\beta}\in B} D_{s_n}(\boldsymbol{\beta}) - D_E(\boldsymbol{\beta}) = o_p(1)$.

A2: Proof of Theorem 2

For simplicity, we denote $\boldsymbol{\beta}(\boldsymbol{\theta}) = \boldsymbol{\beta}$ and $\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}}) = \widehat{\boldsymbol{\beta}}$. Note that

$$\widehat{\boldsymbol{\beta}}_{s_n} = \arg\max_{\boldsymbol{\theta}} D_{s_n}(\boldsymbol{\beta}).$$

Define

$$\begin{aligned}
\mathbf{G}_n(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\theta}} D_{s_n}(\boldsymbol{\beta}) \\
&= \frac{1}{\prod_{j=1}^{M} n_j} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \cdots \sum_{i_M=1}^{n_M} \frac{\partial}{\partial \boldsymbol{\theta}} \left[s_n(\boldsymbol{\beta}^T(\mathbf{X}_{Mi_M} - \mathbf{X}_{(M-1)i_{(M-1)}})) \cdots s_n(\boldsymbol{\beta}^T(\mathbf{X}_{2i_2} - \mathbf{X}_{1i_1})) \right] \\
&= \frac{1}{N} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \cdots \sum_{i_M=1}^{n_M} \Psi(\mathbf{X}_{1i_1}, \mathbf{X}_{2i_2}, \cdots, \mathbf{X}_{Mi_M}; \boldsymbol{\beta})
\end{aligned}$$

where

$$\begin{aligned}
\Psi(\mathbf{X}_{1i_1}, \mathbf{X}_{2i_2}, \cdots, \mathbf{X}_{Mi_M}; \boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\theta}} \left[s_n(\boldsymbol{\beta}^T(\mathbf{X}_{Mi_M} - \mathbf{X}_{(M-1)i_{(M-1)}})) \cdots s_n(\boldsymbol{\beta}^T(\mathbf{X}_{2i_2} - \mathbf{X}_{1i_1})) \right], \\
&= \frac{\partial}{\partial \boldsymbol{\theta}} \left[\prod_{j=1}^{M-1} s_n(\boldsymbol{\beta}^T \mathbf{Z}_{i_{(j+1)}i_j}) \right] \\
&= \sum_{l=1}^{M-1} \left[\prod_{j=1}^{M-1} s_n(\boldsymbol{\beta}^T \mathbf{Z}_{i_{(j+1)}i_j}) \right] \left(1 - s_n(\boldsymbol{\beta}^T \mathbf{Z}_{i_{(l+1)}i_l}) \right) \mathbf{Z}^{(-d)}_{i_{(l+1)}i_l} \\
&= \kappa_n(\mathbf{X}_{1i_1}, \mathbf{X}_{2i_2}, \cdots, \mathbf{X}_{Mi_M}; \boldsymbol{\beta}) \sum_{l=1}^{M-1} \left(1 - s_n(\boldsymbol{\beta}^T \mathbf{Z}_{i_{(l+1)}i_l}) \right) \mathbf{Z}^{(-d)}_{i_{(l+1)}i_l}
\end{aligned}$$

with $\mathbf{Z}_{i_{j+1}i_j} = \mathbf{X}_{(j+1)i_{j+1}} - \mathbf{X}_{ji_j}$, $\mathbf{Z}^{(-d)} = (Z_1, \cdots, Z_{d-1})^T$ and $N = \prod_{j=1}^{M} n_j$. By definition of $\widehat{\boldsymbol{\beta}}_{s_n}$,

$$\mathbf{G}_n(\widehat{\boldsymbol{\beta}}_{s_n}) = \mathbf{0},$$

and $\boldsymbol{\beta}_0$ is such that

$$E(\Psi(\mathbf{X}_{1i_1}, \mathbf{X}_{2i_2}, \cdots, \mathbf{X}_{Mi_M}; \boldsymbol{\beta}_0)) = \mathbf{0}.$$

Since $\mathbf{G}_n(\boldsymbol{\beta})$ is a differentiable function, and $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{s_n} - \boldsymbol{\theta}_0) = o_p(1)$ (result from Theorem 1), hence using Taylor's series expansion we can write

$$\mathbf{0} = \mathbf{G}_n(\widehat{\boldsymbol{\beta}}_{s_n}) = \mathbf{G}_n(\boldsymbol{\beta}_0) + \mathbf{G}_n^{'}(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\theta}}_{s_n} - \boldsymbol{\theta}_0) + \mathbf{R}_n$$

where $\mathbf{G}_n^{'}(\boldsymbol{\beta}_0) = \dfrac{\partial}{\partial \boldsymbol{\theta}^T} \mathbf{G}_n(\boldsymbol{\beta}) \mid_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$ is a $d \times d$ matrix.

Assuming (A4), we can write

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{s_n} - \boldsymbol{\theta}_0) = \left[-\mathbf{G}_n^{'}(\boldsymbol{\beta}_0) \right]^{-1} \sqrt{n} \mathbf{G}_n(\boldsymbol{\beta}_0) + \left[\mathbf{G}_n^{'}(\boldsymbol{\beta}_0) \right]^{-1} \sqrt{n} \mathbf{R}_n. \quad (8)$$

Note that following Theorem 1 where we have $(\widehat{\boldsymbol{\theta}}_{s_n} - \boldsymbol{\theta}_0) = \mathbf{o}_p(1)$, we can write

$$\sqrt{n}\mathbf{R}_n \xrightarrow{p} \mathbf{0}.$$

Following the large sample distribution of multivariate U-statistic (see (25)), it can be shown that

$$\sqrt{n}\mathbf{G}_n(\boldsymbol{\beta}_0) \xrightarrow{d} N_{d-1}\left(\mathbf{0}, \mathbf{B}(\boldsymbol{\beta}_0)\right)$$

where

$$\begin{aligned}
&\mathbf{B}(\boldsymbol{\beta}_0) = \sum_{m=1}^{M} \rho_m^2 \Sigma_{\psi_m}, \\
&\Sigma_{\psi_m} = Var(\tilde{\Psi}_{m1}(\mathbf{X}_{m1})), \\
&\tilde{\Psi}_{m1}(\mathbf{X}_{m1}) = E(\Psi(\mathbf{X}_{11}, \mathbf{X}_{21}, \cdots, \mathbf{X}_{M1})|\mathbf{X}_{m1}), \\
&\rho_m^2 = \frac{n}{n_m}, \quad m = 1, 2, \cdots, M.
\end{aligned}$$

Similarly, using the weak law of large numbers, it can be shown that

$$-\mathbf{G}_n^{'}(\boldsymbol{\beta}_0) = \frac{1}{N}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\cdots\sum_{i_M=1}^{n_M} -\frac{\partial}{\partial \boldsymbol{\theta}^T}\Psi(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}, \cdots, \mathbf{X}_{i_M}; \boldsymbol{\beta}) \xrightarrow{p} \mathbf{A}(\boldsymbol{\beta}_0)$$

where

$$\mathbf{A}(\boldsymbol{\beta}_0) \;=\; E\left(-\frac{\partial}{\partial \boldsymbol{\theta}^T}\Psi(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}, \cdots, \mathbf{X}_{i_M}; \boldsymbol{\beta})\mid_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right).$$

Using Slustky's theorem in equation (8), we can write

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N_{d-1}\left(\mathbf{0}, \Sigma(\boldsymbol{\beta}_0)\right)$$

where $\Sigma(\boldsymbol{\beta}_0) = \mathbf{A}(\boldsymbol{\beta}_0)^{-1}\mathbf{B}(\boldsymbol{\beta}_0)[\mathbf{A}(\boldsymbol{\beta}_0)^{-1}]^T$, known as sandwich variance formula.

Explicit form of the first derivative of $\Psi(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}, \cdots, \mathbf{X}_{i_M}; \boldsymbol{\beta})$ is given as follows:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}^T}\Psi(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}, \cdots, \mathbf{X}_{i_M}; \boldsymbol{\beta}) \;&=\; \frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\left[\prod_{j=1}^{M-1} s_n(\boldsymbol{\beta}^T\mathbf{Z}_{i_{(j+1)}i_j})\right] \\
&=\; \left(\left(\frac{\partial^2}{\partial\theta_u\partial\theta_v}\left[\prod_{j=1}^{M-1} s_n(\boldsymbol{\beta}^T\mathbf{Z}_{i_{(j+1)}i_j})\right]\right)\right), \qquad u, v = 1, 2, \cdots, d-1,
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial^2}{\partial\theta_u\partial\theta_v}\left[\prod_{j=1}^{M-1} s_n(\boldsymbol{\beta}^T\mathbf{Z}_{i_{(j+1)}i_j})\right] \;&=\; \sum_{l=1}^{M-1}\kappa_n(\boldsymbol{\beta})\delta_{n;v}(\boldsymbol{\beta})\left(1 - s_n(\boldsymbol{\beta}^T\mathbf{Z}_{i_{(l+1)}i_l})\right)Z_{i_{(l+1)}i_l;u} - \\
&\quad \sum_{l=1}^{M-1}\kappa_n(\boldsymbol{\beta})s_n(\boldsymbol{\beta}^T\mathbf{Z}_{i_{(l+1)}i_l})\left(1 - s_n(\boldsymbol{\beta}^T\mathbf{Z}_{i_{(l+1)}i_l})\right)Z_{i_{(l+1)}i_l;u}Z_{i_{(l+1)}i_l;v},
\end{aligned}$$

$$\kappa_n(\boldsymbol{\beta}) = \prod_{j=1}^{M-1} s_n(\boldsymbol{\beta}^T \mathbf{Z}_{i_{(j+1)}i_j})$$

and

$$\delta_{n;v}(\boldsymbol{\beta}) = \sum_{k=1}^{M-1} \left(1 - s_n(\boldsymbol{\beta}^T \mathbf{Z}_{i_{(k+1)}i_k})\right) Z_{i_{(k+1)i_k;v}}$$

$$\begin{aligned}
& \Psi(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}, \cdots, \mathbf{X}_{i_M}; \boldsymbol{\beta}) \Psi(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}, \cdots, \mathbf{X}_{i_M}; \boldsymbol{\beta})^T \\
= & \left[\kappa_n(\boldsymbol{\beta}) \sum_{l=1}^{M-1} \left(1 - s_n(\boldsymbol{\beta}^T \mathbf{Z}_{i_{(l+1)}i_l})\right) \mathbf{Z}^{(-d)}_{i_{(l+1)}i_l}\right] \left[\kappa_n(\boldsymbol{\beta}) \sum_{l=1}^{M-1} \left(1 - s_n(\boldsymbol{\beta}^T \mathbf{Z}_{i_{(l+1)}i_l})\right) \mathbf{Z}^{(-d)}_{i_{(l+1)}i_l}\right]^T \\
= & \kappa_n(\boldsymbol{\beta})^2 \sum_{l=1}^{M-1} \sum_{k=1}^{M-1} \left(1 - s_n(\boldsymbol{\beta}^T \mathbf{Z}_{i_{(l+1)}i_l})\right) \left(1 - s_n(\boldsymbol{\beta}^T \mathbf{Z}_{i_{(k+1)}i_k})\right) \mathbf{Z}^{(-d)}_{i_{(l+1)}i_l} \mathbf{Z}^{(-d)T}_{i_{(k+1)}i_k}.
\end{aligned}$$