PREPRINT

國立臺灣大學 數學系 預印本 Department of Mathematics, National Taiwan University

Optimal Receiver Operating Characteristic Manifolds

Yun-Jhong Wu and Chin-Tsang Chiang

November 28, 2011

Optimal Receiver Operating Characteristic Manifolds

Yun-Jhong Wu and Chin-Tsang Chiang
Department of Mathematics, National Taiwan University

November 28, 2011

Abstract

To evaluate overall discrimination capacity of a marker for multi-class classification tasks, the performance function is a natural assessment tool and fully provides the essential ingredients in receiver operating characteristic (ROC) analysis. The optimal ROC manifolds supply a geometric characterization of the magnitude of separation among multiple classes. It has been shown from our foregoing work that the hypervolume under the optimal ROC manifold (HUM) is a well-defined and meaningful accuracy measure only in suitable ROC subspaces. In this article, we provided a rigorous proof for the equality of HUM and its alternative form, the correctness probability, which is directly related to an explicit U-estimator. In addition, extensive simulations are conducted to investigate the finite sample properties of the proposed estimators and the related inference procedures. Further, a rule of thumb is given in application to assess for the HUM. Conclusively, our theoretical framework allows more sophisticated modeling on performance of markers and helps practitioners examine the optimality of applied classification procedures.

Keywords: Gaussian process; Hypervolume; Manifold; Optimal classification; Receiver operating characteristic; U-estimator; Utility

1 Introduction

The past decade has seen the rapid development of multi-classification in various areas of science. For instance, distinguishing species in biology, image recognition in electronic engineering, and pricing strategy in business can be formulated as multi-classification problems. Recently, biomedical researchers have shown an increased interest in accurately determining types of specific diseases or staging cancers, provided that markers contain only limited information regarding the true types. Despite well-established statistical methods for binary classification

problems, such as for distinguishing between diseased and non-diseased patients, it is still questionable whether the existing methodology can appropriately help working scientists to compare performances of various markers, and, if possible, find an optimal marker based on some rational criteria.

A typical task of multi-classification is mainly based on data of the type $(G, \mathbf{Y})$ and a classifier $\widehat{G}$, where $G \in \{1, \ldots, K\}$ represents the true class, $\mathbf{Y} \in \mathcal{Y}$ denotes a univariate or multivariate marker, and $\widehat{G}$ is a random function from $\mathcal{Y}$ to the support of G. The performance probabilities $p_{jk}(\widehat{G}) = P(\widehat{G}(\mathbf{Y}) = j | G = k)$, $j, k \in \{1, \ldots, K\}$, of $(\widehat{G}, \mathbf{Y})$ are commonly used to assess the considered classification procedure. For the sake of numerical stability in estimation, $p_{jk}(\widehat{G})$'s are frequently applied and more preferred than the prediction probabilities $P(G = k | \widehat{G}(\mathbf{Y}) = j)$'s. Furthermore, we can show that there is little connection between the optimization in terms of prediction and performance probabilities although these two are equivalent in binary classification. Since an assessment measure based on performance probabilities is of the form $f(\widehat{G}, \mathbf{Y})$, performance of certain classifiers only represent partial information on discrimination capacity of markers. Thus, evaluation of markers with respective to only a part of classifiers could be too naive to be used to make a fair comparison among markers. Indeed, a rational assessment index of each marker should be a function only of $\mathbf{Y}$ and then is unchanging with chosen classifiers. One of the most practical ways to adopt for this reason is to address the marker of interest with respect to the overall performance of all classifiers. To achieve this research aim, receiver operating characteristic (ROC) analysis, a technique initially only for binary classification tasks, has been extended to multi-classification in recent years.

Meaning of optimality in classification could be various, but some optimal criteria are shown to be equivalent in the sense of overall performance. A seminal work on optimal ROC manifolds was promulgated by Scurfield (1998) via maximizing $\sum_{k=1}^{K} P(\widehat{G} = k, G = k)$. The author constructed ROC manifolds based on performance probabilities of optimal deterministic clas-

sifiers with the maximal sum of true probabilities and derived that HUM equals to correctness probability for the ternary classification procedures; the optimal classifiers can be represented as combinations of linear classifiers in the decision space spanned by log-likelihood ratio scores. Since K^2 performance probabilities are available to describe a K-classification procedure, Edwards et al. (2004) explained that it would be insufficient to describe the complete information of a classifier only based on true probabilities $p_{kk}(\widehat{G})$'s for $K \geq 3$; they further suggested to maximize the expected utility (or minimize the expected cost), and this criterion can be formulated in a manner of linear classifiers in the decision space. For ternary classification, He and Frey (2006) indicated that the utility classifiers have maximal sum of true probabilities under the setting of equal error utilities. As an alternative approach, Schubert et al. (2011) utilized Minkowski's functionals to determine the optimal classifier. Roughly speaking, the functional is to define an optimal classifier as that with the minimal misclassification rates under constraints on ratios among these probabilities. The criterion is essentially analogue to maximizing the expected utility under some conditions, although their illustrated examples do not actually achieve its optimality and it is difficult to give a feasible formula of their defined optimal classifier. Due to the difficulty in visualization for performance in multi-classification, a vital issue arises to define an appropriate summary index for the performance of a marker. Naturally, HUM is a direct extension of the area under the ROC curve (AUC) and was employed in many foregoing works. In binary classification, the induced optimal ROC curve certainly separates the ROC space into two regions. However, optimal ROC manifolds may be unable to enclose a bounded set and, hence, the well-definition of HUM might be thrown in doubt. Besides, Edwards et al. (2005) used some examples to explain that both of the resulting HUMs from near-perfect and non-informative markers are near zero; these authors further concluded that HUM is not a suitable summary index for performance of a marker.

The breakthrough results we have achieved are initially based on a theoretical formulation in

terms of utility, which concisely describes current results regarding multi-class ROC analysis. The groundwork leads a better understanding of some geometric characteristics of optimal ROC manifolds, such as regularity (see Jost, 2008) and smoothness, whereas non-optimal ROC manifolds may not enjoy these attractive features; moreover, the asymptotic process of empirical optimal ROC manifolds is established. We also address the sufficient and necessary conditions to ensure the well-behaved HUM; one can clearly interpret some peculiar numerical and algebraic results occurring in the foregoing works and further advocate practitioners to create a handy summary assessment. Instead of the setting in ternary classification, we borrowed a tool in graph theory to confirm the validation of $HUM = CP$ for any K-class optimal procedures. Thus, by using this explicit and meaningful probability expression, an U-estimation for HUM then becomes applicable for more general classification procedures. In considering practical implications, we proposed an estimation approach for HUM with related inference procedures through some widely-used models on the relationship between G and $\mathbf{Y}$ through a prospective or retrospective perspective. Furthermore, an empirical rule based on partial-classification HUMs is proposed to assist practitioners to evaluate the disciminability. Although our work focuses on continuous markers, most of these results are comfortably adapted to evaluation of discrete or mixture markers and could serve as a basis for more sophisticated statistical methods.

On the whole, based on the properties of optimal ROC manifolds we have established, we provide estimation and inference procedure for discriminability of multi-classification markers. The properties enables us to draw pointwise and functional inference for optimal ROC manifolds in Section 2. Section 3 is devoted to estimation and model-based inference procedures for HUM. Numerical experiments and an application to empirical data in Section 4 illustrate the practicality of our developed methodology. Finally, Section 6 summarizes the findings in this study and make some remarks for future research.

2 Optimal ROC Manifolds

Some researchers have worked on construction of ROC manifolds; however, without optimality in classifiers, the so-called ROC manifold could be an arbitrary subset of a projection of the performance set

$$\phi(\mathcal{C}) = \{\phi(\widehat{G}) : \widehat{G} \in \mathcal{C}\},$$

where $\phi(\widehat{G})$ is the performance of $\widehat{G}$ defined as $\text{vec}[p_{jk}(\widehat{G})]$, in the ROC space

$$\mathcal{R} = \{\mathbf{p} = \text{vec}[p_{jk}] : \sum_{j=1}^{K} p_{jk} = 1 \ \forall k = 1, \ldots, K\}.$$

rather than a manifold in the context of geometry. Therefore, few features of the ROC manifold sets could be pinpointed, and estimation of ROC manifold sets and related summary measures might lead to a more complicated situation. We hence introduce optimal ROC manifolds for multi-classification as an extension of optimal ROC curves for binary classification. For K-classification tasks, there are K redundant coordinates in $\mathcal{R}$. Practically, not all K^2 performance probabilities are of interest. We can further consider a ROC subspace $\mathcal{R}_S$ where S denotes the set of coordinates of concern. In the sequel, sets or operators with the subscripted S denotes that they are restricted in the ROC subspace $\mathcal{R}_S$

Indeed, the performance set $\phi(\mathcal{C})$ is a convex and compact set and hence can be completely characterized through investigating its boundary set $\partial\phi(\mathcal{C})$; these features also hold in arbitrary $\mathcal{R}_S$. Moreover, $\partial\phi(\mathcal{C})$ is also able to be regarded as a function only depending on $\mathcal{Y}$. Through the connection of admissibility and maximizing-utility criterion, we have obtained severals results: First, performance of each admissible classifier is located in $\partial\phi(\mathcal{C})$; this justifies using the optimal ROC manifold

$$M_S = \{\phi_S(\widehat{G}) : \widehat{G} \text{ is admissible in } S.\}$$

as a measure of discriminability of markers. Second, maximizing-utility criterion gives a natural parametric system to describe M_S. Hence, M_S can be treated as actually a function of $\mathbf{u} \in \mathcal{U}$

on an interesting set $\mathcal{U}$ of utility values, denoted by $M_S(\mathbf{u})$, with some desirable functional asymptotic properties of the corresponding empirical estimators. For a specific p_{jk}, since in the perspective of geometry points in M_S can be regarded as the performance with the highest p_{jk} for every fixed performance probabilities $S\backslash\{p_{jk}\}$, $M_S(\mathbf{u})$ can be rewritten as $M_S^*(\mathbf{p}_{S\backslash\{p_{jk}\}})$.

An admissible classifier can be derived as that $\widehat{G}(\mathbf{y}) = k$ if

$$L(\mathbf{y}) \in D_k(\mathbf{u}) = \bigcap_{j\neq k}\{L(\mathbf{y}) : \sum_{i=1}^{K}(u_{ki} - u_{ji})L_{iK}(\mathbf{y}) \geq 0, \mathbf{y} \in \mathcal{Y}\}, \tag{1}$$

where $L(\mathbf{y}) = (L_{1K}(\mathbf{y}), \ldots, L_{(K-1)K}(\mathbf{y}))^\top$ with the likelihood ratio L_{ij} between ith and jth classes. We should note that each $D_k(\mathbf{u})$ is an intersection of $K-1$ half spaces in the space spanned by likelihood ratio scores and so simple enough to be practically applied.

Based on a random sample $\{(\mathbf{Y}_m, G_m)\}_{m=1}^n$, an empirical estimator of $p_{jk}(\widehat{G}_\mathbf{u})$ is proposed to be

$$\widehat{p}_{jk}(\widehat{G}_\mathbf{u}) = \widehat{n}_k^{-1}\sum_{m=1}^{n} 1(L(\mathbf{Y}_m) \in D_j(\mathbf{u}))1(G_m = k),$$

where $\widehat{n}_k = \sum_{m=1}^n 1(G_m = k)$ and $\widehat{M}_S(\mathbf{u})$ is set to be $\widehat{\phi}_S(\widehat{G}_\mathbf{u})$ with $\widehat{p}_{jk}(\widehat{G}_\mathbf{u})$ substituting for $p_{jk}(\widehat{G}_\mathbf{u})$. Given fixed $p_{j'k'} = p^*_{j'k'}$ for all $p_{j'k'} \in S\backslash\{p_{jk}\}$, the asymptotic normality of $\widehat{M}_S(\mathbf{u})$ enables us to have an approximate $1-\alpha$, $0 < \alpha < 1$, confidence interval for $M_S^*(\mathbf{p}_{S\backslash\{p_{jk}\}})$. By replacing $p_{jk}(\widehat{G}_\mathbf{u})$ with $\widehat{p}_{jk}(\widehat{G}_\mathbf{u})$, $M_S^*(\mathbf{p}_{S\backslash\{p_{jk}\}})$ and its asymptotic variance are naturally estimated by

$$\widehat{M}_S^*(\mathbf{p}_{S\backslash\{p_{jk}\}}) = \max\{\widehat{p}_{jk}(\widehat{G}_\mathbf{u}) : \widehat{p}_{j'k'}(\widehat{G}_\mathbf{u}) \geq p^*_{j'k'}\ \forall p_{j'k'} \in S\backslash\{p_{jk}\}\} \tag{2}$$

and

$$\widehat{\sigma}^2(\widehat{M}_S^*(\mathbf{p}_{S\backslash\{p_{jk}\}})) = \widehat{M}_S^*(\mathbf{p}_{S\backslash\{p_{jk}\}})(1 - \widehat{M}_S^*(\mathbf{p}_{S\backslash\{p_{jk}\}})).$$

Thus, the normal-type confidence interval for $M_S^*(\mathbf{p}_{S\backslash\{p_{jk}\}})$ can be constructed by

$$\widehat{M}_S^*(\mathbf{p}_{S\backslash\{p_{jk}\}}) \pm z_{1-\alpha/2}n^{-1/2}\widehat{\sigma}\widehat{M}_S^*(\mathbf{p}_{S\backslash\{p_{jk}\}}), \tag{3}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of a standard normal distribution. The interval in (3) provides us a base to build up pointwise confidence bands of $p_{jk}(\widehat{G}_\mathbf{u})$. Similar construction can

also be done for performance probabilities of any non-optimal classifier. Although it could be harder to be visualized and interpreted, a pointwise confidence region of $\{p_{jk}(\widehat{G}_{\mathbf{u}}) : \mathbf{u} \in \mathcal{U}\}$ can be constructed in a similar manner. By applying the central limit theorem and the Slutsky's theorem, one can readily derive that

$$n^{1/2}(\widehat{M}_S(\mathbf{u}) - M_S(\mathbf{u})) \xrightarrow{d} N_{\#S}(0, \mathbf{\Sigma}(\mathbf{u})), \tag{4}$$

where the asymptotic covariance between $\widehat{p}_{jk}(\widehat{G}_{\mathbf{u}})$ and $\widehat{p}_{j'k'}(\widehat{G}_{\mathbf{u}})$ is

$$\mathbf{\Sigma}_{jk,j'k'}(\mathbf{u}) = \begin{cases} p_{jk}(\widehat{G}_{\mathbf{u}})(1 - p_{jk}(\widehat{G}_{\mathbf{u}})) & \text{for } (j,k) = (j',k'), \\ -p_{jk}(\widehat{G}_{\mathbf{u}})p_{j'k'}(\widehat{G}_{\mathbf{u}}) & \text{for } j \neq j', k = k', \\ 0 & \text{otherwise.} \end{cases}$$

An approximate $1-\alpha$ pointwise confidence region is then given by

$$\{\mathbf{p}_S : n(\mathbf{p}_S - \widehat{M}_S(\mathbf{u}))^T \widehat{\mathbf{\Sigma}}^{-1}(\mathbf{u})(\mathbf{p}_S - \widehat{M}_S(\mathbf{u})) < \chi^2_{\#S,\alpha}\}, \tag{5}$$

where $\chi^2_{\#S,\alpha}$ is the αth quantile of a chi-square distribution with the degrees of freedom $\#S$.

Before establishing a simultaneous confidence region for M_S, the related asymptotic properties of $\widehat{\phi}_S(\widehat{G})$ for some $\widehat{G} \in \mathcal{C}$ are derived first. The main difficulty in theoretical development is that classifiers of interest could be a family including perplexing functions. In contrast, an optimal classifier $\widehat{G}_{\mathbf{u}}$ can be reformulated as a linear combination of the indicator functions $\sum_{k=1}^{K} k1(L(\mathbf{Y}) \in D_k(\mathbf{u}))$ on $\mathcal{D}$ with $D_k(\mathbf{u})$ begin an intersection of $K-1$ half-spaces. It is well-known that the collection of half-spaces of $(K-1)$ dimensional space is of the Vapnik-Chervonenkis (VC) dimension K. Thus, the collection of $D_k(\mathbf{u})$, a subset of intersections of K half-spaces indexed by $\mathbf{u}$, also has finite VC-dimension. That is, $D_k(\mathbf{u})$ and all optimal classifiers are of VC-class. This ensures that the covering number of the collection of $\widehat{G}_{\mathbf{u}}$ grows in a polynomial way. By Theorem 2.6.4 in Vaart and Wellner (1996), the finite VC dimension of $\{\widehat{G}_{\mathbf{u}}\}_{\mathbf{u}\in\mathcal{U}}$ gives a universal bound of the covering number of optimal classifiers. Together with the measurability of $\widehat{G}_{\mathbf{u}}$ and Theorems 2.4.3 and 2.5.2 in the same reference, the family of optimal classifiers are of P-Glivenko-Cantelli and also of P-Donsker class. The results are summarized by the following theorem.

Theorem 2.1. *The empirical estimator of an optimal ROC manifold* $\widehat{M}_S(\mathbf{u})$ *has the functional asymptotic properties:*

i. $\sup_{\mathbf{u}\in\mathcal{U}} \|\widehat{M}_S(\mathbf{u}) - M_S(\mathbf{u})\| \xrightarrow{a.s.} 0.$

ii. $n^{1/2}(\widehat{M}_S(\mathbf{u}) - M_S(\mathbf{u}))$ *converges in distribution to a Gaussian process with mean zero and a covariance function, where the asymptotic covariance between* $\widehat{p}_{jk}(\widehat{G}_{\mathbf{u}_1})$ *and* $\widehat{p}_{j'k'}(\widehat{G}_{\mathbf{u}_2})$*, say* $\mathbf{\Sigma}_{jk,j'k'}(\mathbf{u}_1, \mathbf{u}_2)$*, is*

$$P(L(\mathbf{Y}) \in D_j(\mathbf{u}_1) \cap D_{j'}(\mathbf{u}_2) | G = k) - p_{jk}(\widehat{G}_{\mathbf{u}_1}) p_{j'k}(\widehat{G}_{\mathbf{u}_2}),$$

for $k = k'$ *and* 0 *otherwise.*

With Theorem 2.1, a simultaneous confidence region for $\{M_S(\mathbf{u}) : \mathbf{u} \in \mathcal{U}\}$ can be built up through

$$P(n \sup_{\mathbf{u}\in\mathcal{U}} (\widehat{M}_S(\mathbf{u}) - M_S(\mathbf{u}))^\top \mathbf{\Sigma}^{-1}(\mathbf{u})(\widehat{M}_S(\mathbf{u}) - M_S(\mathbf{u})) < c_\alpha) = 1 - \alpha,$$

whereas the quantile value c_α is not easy to obtain directly. By using

$$n^{1/2}(\widehat{M}_S(\mathbf{u}) - M_S(\mathbf{u})) = n^{-1/2} \sum_{m=1}^{n} \mathbf{N}_m(\mathbf{u}) + r_n(\mathbf{u})$$

with $\mathbf{N}_m(\mathbf{u}) = \mathrm{E}[\widehat{M}_S(\mathbf{u}) - M_S(\mathbf{u}) | \mathbf{Y}_m]$ and $\sup_{\mathbf{u}\in\mathcal{U}} \|r_n(\mathbf{u})\| = o_p(1)$, one can follow the same argument in Lin et al. (2000) to show that

$$P_n(n^{-1} \sup_{\mathbf{u}\in\mathcal{U}} (\sum_{m=1}^{n} \widehat{\mathbf{N}}_m(\mathbf{u}) w_m)^\top \widehat{\mathbf{\Sigma}}^{-1}(\mathbf{u}) (\sum_{m=1}^{n} \widehat{\mathbf{N}}_m(\mathbf{u}) w_m) < c_\alpha) \xrightarrow{p} 1 - \alpha, \tag{6}$$

where P_n is the probability measure conditioning on $\{(G_m, \mathbf{Y}_m)\}_{m=1}^n$, w_m's are independently drawn from a standard normal distribution, $\widehat{\mathbf{\Sigma}}(\mathbf{u})$ is a moment estimator of $\mathbf{\Sigma}(\mathbf{u})$ in (4), and $\widehat{\mathbf{N}}_m(\mathbf{u})$ is a consistent estimator of $\mathbf{N}_m$ with

$$\widehat{N}_{m,jk}(\mathbf{u}) = 1(L(\mathbf{Y}_m) \in D_j(\mathbf{u})) 1(G_m = k) - \widehat{p}_{jk}(\widehat{G}_{\mathbf{u}}).$$

The convergence property in (6) enables us to estimate c_α by the Monte-Carlo quantile c^*_α. An approximate $1 - \alpha$ simultaneous confidence region for $\{M_S(\mathbf{u}) : \mathbf{u} \in \mathcal{U}\}$ is then given by

$$\{\mathbf{p}_S : n(\mathbf{p}_S - \widehat{M}_S(\mathbf{u}))^\top \widehat{\mathbf{\Sigma}}^{-1}(\mathbf{u})(\mathbf{p}_S - \widehat{M}_S(\mathbf{u})) < c^*_\alpha, \forall \mathbf{u} \in \mathcal{U}\}. \tag{7}$$

In high-dimensional $\mathcal{R}_S$, it is usually difficult to concretely illustrate implications from the constructed confidence regions (5) and (7) for optimal ROC manifolds. Researchers might like to adopt the coverage probability of a confidence region to evaluate its performance or draw inferences on some meaningful summary indices of M_S.

3 Hypervolumes under Optimal ROC Manifolds

Once there are more than three classes and the number of considered p_{jk}'s is greater than three, the ROC subspace of interest might involve complication in visualization. An appropriate summary index of the performance of a marker becomes practically important. Ideally, an applied summary index should satisfy at least two requirements: the index must facilitate comparisons for all markers so that their performances are comparable with each other; the index should provide a reasonable ordering of performance of markers. For binary classification tasks, AUC is the most widely-used accuracy measure, and the measure could be well-defined since the corresponding ROC curve can usually separate a ROC subspace. The accuracy measure HUM is a natural extension of AUC and has been proposed and applied in former literature, whereas its features still remain largely unexplored. Since the optimality of ROC manifolds guarantees its continuity, it is possible to constitute a separation of a ROC subspace by the manifolds. For this reason, we use the term HUM, denoted by V_S, referring hypervolume under M_S until further notice.

For binary classification, the dimensionality of $\phi(\mathcal{C})$ equals 2 and then ensures the existence of optimal AUC. However, even with optimality, HUM is generally not well-defined or might be ill-behaved for $K \geq 3$. A series of results are further established in the following to clearly characterize such an accuracy assessment. Indeed, the set under the opimal ROC manifolds M_S has a well-defined volume V_S form if and only if the manifolds is restricted in $\mathcal{R}_S$ for either

$S = \{p_{kk} : k = 1, \ldots, K\}$ or $S = \{p_{k\sigma(k)} : k = 1, \ldots, K, k \neq \sigma(k)\}$. In this situation, the HUM as a summary index actually assist us to evaluate the discriminability of markers. Furthermore, optimal classifiers can be simplified as that $\mathbf{Y}$ is classified as the kth class if $\mathbf{Y} \in D_k(\mathbf{u})$ with

$$D_k(\mathbf{u}) = \bigcap_{j \neq k} \{L(\mathbf{y}) : L_{jk}(\mathbf{y}) \geq \frac{u_{kk}}{u_{jj}}, \mathbf{y} \in \mathcal{Y}\}.$$

Utilities u_{jk}'s are also able to be reparametrized as threshold values $c_k = u_{KK}/u_{kk}$, $k = 1, \ldots, K-1$, and, for convenience, let $\mathbf{c} = (c_1, \ldots, c_{K-1})$.

3.1 Estimation and Inference Procedures for HUM

An estimation for HUM directly through its Riemann integral usually involves a computational difficulty in a high dimensional space. A related algorithm might lead to numerical instability and terrible computational cost even for $K = 3$. A probability expression of HUM would greatly enhance the efficiency of its estimation. Indeed, the HUM corresponding to $S = \{p_{k\sigma(k)}\}_{k=1}^{K}$ has an explicit connection with the event that the K subjects $(G = k, \mathbf{Y}_k)$'s can be classified to the class $\sigma(k)$ for each k by one classifier $\widehat{G}_{\mathbf{u}}$. To simplify the succeeding discussion, we define

$$h_S(\mathbf{Y}_1, \ldots, \mathbf{Y}_K) = 1(\exists\ \mathbf{u} : \widehat{G}_{\mathbf{u}}(\mathbf{Y}_1) = \sigma(1), \ldots, \widehat{G}_{\mathbf{u}}(\mathbf{Y}_K) = \sigma(K)). \tag{8}$$

The expectation of $h_S(\mathbf{Y}_1, \ldots, \mathbf{Y}_K)$ given $\{G_1 = 1, \ldots, G_K = K\}$ is the correctness probability. For ternary classification, Scurfield (1998) has shown $HUM = CP$, which is convenient to access an U-estimator for the probability expression of the HUM. However, further confirming the equality for any $K \geq 4$ involves more tedious calculations because it needs to compute integrals on some domain like the form $\partial D_K(\mathbf{u}) \times \{\times_{k=1}^{K-1} D_k(\mathbf{u})\}$. An alternative exposition in the sense of spanning trees in graph theory can be utilized to establish the validity of a sharpened version of Scurfield's theorem. The following technical lemma is devoted to the main theorem, but is of separated interest that an optimal classifier gives the maximum likelihood prediction.

Lemma 3.1. *Given any permutation σ_0 associated with S,*

$$h_S(\mathbf{Y}_1, \ldots, \mathbf{Y}_K) = 1(\prod_{k=1}^{K} f_k(\mathbf{Y}_{\sigma_0(k)}) \geq \prod_{k=1}^{K} f_k(\mathbf{Y}_{\sigma(k)}) \ \forall \sigma).$$

Proof. See Appendix. □

Lemma 3.1 makes an equivalent statement of the event that $\{\mathbf{Y}^{(k)}\}_{k=1}^K$ can be correctly classified by one optimal classifier. Particularly, for σ_0 being the identity permutation, the optimal classifiers guarantee that probabilities of subjects being classified into the true classes are higher than into other classes. We can now rephrase $HUM = CP$ for any K-classification as follows:

Theorem 3.2. *Suppose that $S = \{p_{11}, \ldots, p_{KK}\}$ or $S = \{p_{k\sigma(k)} : \sigma(k) \neq k\}_{k=1}^K$ for some σ. Then,*

$$V_S = \mathrm{E}[h_S(\mathbf{Y}_1, \ldots, \mathbf{Y}_K) | G_1 = 1, \ldots, G_K = K\}].$$

Proof. See Appendix. □

Lemma 3.1 and Theorem 3.2 justify a more general U-estimator of V_S as

$$\widehat{V}_S = \frac{\sum_{\{i_1,\ldots,i_K\}\subset[n]} h_S(\mathbf{Y}_{i_1}, \ldots, \mathbf{Y}_{i_K}) \prod_{k=1}^K 1(G_{i_k} = k)}{\sum_{\{i_1,\ldots,i_K\}\subset[n]} \prod_{k=1}^K 1(G_{i_k} = k)}. \tag{9}$$

Following the asymptotic normality of U-statistics in the former works, one can also show that

$$n^{1/2}(\widehat{V}_S - V_S) \xrightarrow{d} N(0, \xi), \tag{10}$$

where $\xi = \sum_{k=1}^K p_k^{-1} \mathrm{Var}(\mathrm{E}[h_S(\mathbf{Y}_1, \ldots, \mathbf{Y}_K) | \mathbf{Y}_k, G_1 = 1, \ldots, G_K = K])$. It is straightforward to obtain a consistent estimator of ξ as

$$\widehat{\xi} = \sum_{k=1}^{K} \widehat{n}_k^{-1} \sum_{m=1}^{n} (\widehat{V}_S(m,k) - \widehat{V}_S)^2 1(G_m = k), \tag{11}$$

where

$$\widehat{V}_S(m,k) = \prod_{k'' \neq k} \widehat{n}_{k''}^{-1} \sum_{\substack{\{i_1,\ldots,i_K\} \subset [n] \\ i_k = m}} h_S(\mathbf{Y}_{i_1}, \ldots, \mathbf{Y}_{i_K}) \prod_{k'=1}^{K} 1(G_{i_{k'}} = k').$$

From (9) and (10), we can readily construct an approximate $1-\alpha$ confidence interval for V_S as

$$\widehat{V}_S \pm n^{-1/2} z_{1-\alpha/2} \widehat{\xi}^{1/2}.$$

3.2 Model-based HUM

In the preceding work, h_S in (9) is known to be a composite function of an unknown likelihood ratio transformation. Appropriate modeling would simplify classification tasks, such as sequential classification procedures in MLR models. Thus, a model-based estimator for V_S usually has a simple representation. However, even with specific parametric models, (10) cannot be obtained straightforward because some parameters in h_S are still unknown. For explanatory simplicity, we only consider $S = \{p_{11}, \ldots, p_{KK}\}$. The inference for $\{p_{k\sigma(k)} : \sigma(k) \neq k\}_{k=1}^{K}$ can be established in a parallel manners.

A direct approach to estimation for V_S is to model the likelihood ratios

$$L_{kK}(\mathbf{y}) = g_k(\mathbf{y}; \boldsymbol{\theta}_0) \tag{12}$$

for $k = 1, \ldots, K-1$, where g_k is a specified function with unknown parameters $\boldsymbol{\theta}_0$. Correspondingly, $L_{jk}(\mathbf{y}) = g_{jk}(\boldsymbol{y}; \boldsymbol{\theta}_0) = g_j(\boldsymbol{y}; \boldsymbol{\theta}_0)/g_k(\boldsymbol{y}; \boldsymbol{\theta}_0)$ and $L(\mathbf{y}) = g(\mathbf{y}; \boldsymbol{\theta}) = (g_1(\mathbf{y}; \boldsymbol{\theta}), \ldots, g_{K-1}(\mathbf{y}; \boldsymbol{\theta}))^\top$. Since $L_{kK}(\mathbf{y})$ is proportional to $P(G = k|\mathbf{Y} = \mathbf{y})/P(G = K|\mathbf{Y} = \mathbf{y})$, the overall accuracy measure, denoted by $V(\boldsymbol{\theta}_0)$, of markers is irrelevant with sampling schemes. Thus, $V(\boldsymbol{\theta}_0)$ can be derived through modeling either f_k's in a retrospective perspective or $P(G = k|\mathbf{Y} = \mathbf{y})$'s in a prospective one. Under the model (12) with given $\boldsymbol{\theta}_0$, the estimated HUM in (9) can be parametrically expressed as

$$\widehat{V}(\boldsymbol{\theta}_0) = \frac{\sum_{\{i_1,\ldots,i_K\}\subset[n]} h_S(\mathbf{Y}_{i_1},\ldots,\mathbf{Y}_{i_K};\boldsymbol{\theta}_0)\prod_{k=1}^{K} 1(G_{i_k}=k)}{\sum_{\{i_1,\ldots,i_K\}\subset[n]}\prod_{k=1}^{K} 1(G_{i_k}=k)} \tag{13}$$

with $h_S(\mathbf{Y}_{i_1},\ldots,\mathbf{Y}_{i_K};\boldsymbol{\theta}_0) = 1(\sum_{k=1}^{K}\ln g_{k\sigma(k)}(\mathbf{Y}_{i_k};\boldsymbol{\theta}_0) \geq 0\ \forall\sigma)$. By replacing $\boldsymbol{\theta}_0$ in (13) with a $\sqrt{n}$-consistent estimator $\widehat{\boldsymbol{\theta}}$, the final estimator $\widehat{V}(\widehat{\boldsymbol{\theta}})$ is naturally obtained. Under some suitable conditions, $\widehat{V}(\widehat{\boldsymbol{\theta}})$ and $\widehat{V}(\boldsymbol{\theta}_0)$ are shown to be asymptotically equivalent. The limiting distribution of $n^{1/2}(\widehat{V}(\widehat{\boldsymbol{\theta}}) - V(\boldsymbol{\theta}_0))$ is further established by the following theorem:

Theorem 3.3. *Under the validity of the model (12) with $g(\mathbf{Y};\boldsymbol{\theta})$ being Lipschitz continuous on a compact set $\boldsymbol{\Theta}$,*

$$n^{1/2}(\widehat{V}(\widehat{\boldsymbol{\theta}}) - V(\boldsymbol{\theta}_0)) \xrightarrow{d} N(0,\xi(\boldsymbol{\theta}_0)),$$

provided $\xi(\boldsymbol{\theta}_0) = \sum_{k=1}^{K} p_k^{-1}\mathrm{Var}(\mathrm{E}[h_S(\mathbf{Y}_1,\ldots,\mathbf{Y}_K;\boldsymbol{\theta}_0)|\mathbf{Y}_k, G_1=1,\ldots,G_K=K]) > 0.$

Proof. See Appendix. □

From Theorem 3.3, an approximate $1-\alpha$ confidence interval for $V(\boldsymbol{\theta}_0)$ is given by

$$\widehat{V}(\widehat{\boldsymbol{\theta}}) \pm n^{-1/2} z_{1-\alpha/2}\widehat{\xi}^{1/2}(\widehat{\boldsymbol{\theta}}), \tag{14}$$

where $\widehat{\xi}(\widehat{\boldsymbol{\theta}})$ is the same as that in (11) with $\widehat{\boldsymbol{\theta}}$ substituting for $\boldsymbol{\theta}_0$. Generally, the Lipschitz continuity of g is assured through some widely applied models. At the end of this discussion, we illustrate three parametric models and concisely rewrite their h_S's in (8).

Example 3.1. Under the validity of a multinomial logistic model, one has $g_k(\mathbf{y};\boldsymbol{\theta}_0) = \exp(\boldsymbol{\theta}_{0k}^{\top}\mathbf{Y})P(G=K)/P(G=k)$, $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{01}^{\top},\ldots,\boldsymbol{\theta}_{0(K-1)}^{\top})^{\top}$. It follows that h_S is with the exposition

$$\prod_{\sigma} 1(\sum_{k=1}^{K}(\boldsymbol{\theta}_{0k}-\boldsymbol{\theta}_{0\sigma(k)})^{\top}\mathbf{Y}_k \geq 0), \boldsymbol{\theta}_{0K} = \mathbf{0}.$$

Example 3.2. When $\mathbf{Y}|G=k$ follows a multivariate normal distribution with mean $\boldsymbol{\mu}_{0k}$ and covariance matrix $\boldsymbol{\Sigma}_{0k}$, $k=1,\ldots,K$, the induced h_S is of the form

$$\prod_{\sigma} 1\Big(\sum_{k=1}^{K}(\mathbf{Y}_k - \boldsymbol{\mu}_{0\sigma(k)})^\top \boldsymbol{\Sigma}_{0\sigma(k)}^{-1}(\mathbf{Y}_k - \boldsymbol{\mu}_{0\sigma(k)}) \geq \sum_{k=1}^{K}(\mathbf{Y}_k - \boldsymbol{\mu}_{0k})^\top \boldsymbol{\Sigma}_{0k}^{-1}(\mathbf{Y}_k - \boldsymbol{\mu}_{0k})\Big).$$

Example 3.3. Let Y be univariate with the corresponding family of distributions f_k's satisfying the MLR condition with respect to $\boldsymbol{\theta}_0$. Specifically, $g_{k_1k_2}(y;\boldsymbol{\theta}_0) = g_{k_1}(y;\theta_{0k_1})/g_{k_2}(y;\theta_{0k_2})$ is strictly increasing in y for $\theta_{0k_1} > \theta_{0k_2}$. Then, h_S can be simplified as

$$\prod_{k_1 \neq k_2} 1(Y_{k_1} > Y_{k_2})1(\theta_{0k_1} > \theta_{0k_2}).$$

4 Numerical Experiments

To investigate the finite sample properties of $\widehat{V}(\widehat{\boldsymbol{\theta}})$ and the proposed inference procedure for $V(\boldsymbol{\theta}_0)$, we conducted a series of numerical experiments with the three models illustrated in Examples 3.1 through 3.3 in the first three sections. The simulation results were based on 2,000 replications of each assignment to preserve values stabilized up to the third decimal digit. For each simulation setting, means, standard deviations (sd), and standard errors (se) of the estimated accuracy measures with known and estimated parameters were exhibited. In addition, the relation between $\widehat{V}(\widehat{\boldsymbol{\theta}})$ and $\widehat{V}(\boldsymbol{\theta}_0)$ was detected under variant sample sizes. As for the assessment of the constructed confidence interval in (14), we provided the quantile intervals (QI) of 2,000 $\widehat{V}(\widehat{\boldsymbol{\theta}})$ with length (L_q), the averages of 0.95 normal-type confidence intervals (CI_z) of $V(\boldsymbol{\theta}_0)$ with length (L_z), and the empirical coverage probabilities (E_{cp}).

4.1 Scenario I: Multinomial Logistic Regression

Assuming optimality of linear combinations of original markers, multinomial logistic regression in Example 3.1 is a widely used approach to the multi-classification problem. Let (Y_1, Y_2, Y_3)

be a trivariate normal distribution with mean zero, standard deviation $(1, 1, 1)$, and correlation coefficient of 0.2, 0.5, or 0.8. The first investigated model was designed to be

$$P(G = k|\mathbf{Y} = \mathbf{y}) = \frac{\exp((1, \mathbf{y}^\top)\boldsymbol{\theta}_{0k})}{1 + \sum_{j=1}^{2} \exp((1, \mathbf{y}^\top)\boldsymbol{\theta}_{0j})} \tag{15}$$

for $k = 1, 2$ with $\boldsymbol{\theta}_{01} = (-0.2, 1, 1, -1)^\top$ and $\boldsymbol{\theta}_{02} = (0.2, 1, -2, 1)^\top$. The induced proportions $(P(G = 1), P(G = 2), P(G = 3))$ for $\rho = 0.2, 0.5$, and 0.8 are further computed to be $(0.327, 0.426, 0.247)$, $(0.320, 0.418, 0262)$, and $(0.269, 0.402, 0.329)$, respectively.

Tables 1 and 2 report the means and standard deviations of $\widehat{V}(\boldsymbol{\theta}_0)$ and $\widehat{V}(\widehat{\boldsymbol{\theta}})$ under different correlations of markers and sample sizes. The numerical results indicate that $\widehat{V}(\widehat{\boldsymbol{\theta}})$ overestimates $V(\boldsymbol{\theta}_0)$ and its asymptotic variance is slightly underestimated in some case. In addition, the difference between $\widehat{V}(\boldsymbol{\theta}_0)$ and $\widehat{V}(\widehat{\boldsymbol{\theta}})$ steadily declines as n increases although it is relatively notable for small sample sizes. The results manifest their asymptotic equivalence for an appropriately large sample size. One can see that the estimates of bounds in normal-type confidence intervals are fairly reliable. Suffering from the small sample size in the case of $n = 150$, the estimated normal-type intervals are notably biased and the empirical coverage probabilities are substantially smaller than the nominal level. Furthermore, it is detected from Table 2 that the empirical coverage rates for $\rho = 0.2$ are lower than those for $\rho = 0.5$ and 0.8. This phenomenon emerges due to a high value of $V(\boldsymbol{\theta}_0)$ and is apparent in the next numerical experiment. Generally, the empirical coverage probabilities are very close to the assigned nominal level.

4.2 Scenario II: Multivariate Normal Marker

In classification, multivariate markers are popularly modeled via multivariate normal distributions. With a common correlation structure, linear classifiers based on original markers actually achieve optimality with

$$h_S(\mathbf{Y}_1, \ldots, \mathbf{Y}_K) = \prod_{\sigma} 1\Big(\sum_{k=1}^{K} (\boldsymbol{\mu}_{0\sigma(k)} - \boldsymbol{\mu}_{0k})^\top \boldsymbol{\Sigma}_0^{-1} \mathbf{Y}_k \geq 0\Big).$$

The distributions of $\mathbf{Y}$ conditioning on $G = 1, 2$, and 3 were specified with means $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, -1)$, respectively, and the common standard deviation $(1, 1, 1)$ and correlation coefficient of 0.2, 0.5, or 0.8. The proportion of each class was further set to be equal on average. To evaluate the impact of over-parameterization, we used pooled and group-specific sample covariance matrices in classification. Both of the procedures are optimal in classification, whereas estimation of redundant parameters usually leads to some numerical instability and loses efficiency.

It is detected in Tables 3 and 4 that the biases rapidly become negligible with increasing sample sizes, although $V(\boldsymbol{\theta}_0)$ tends to be overestimated and an apparent bias appears in over-parameterization. Specifically, the biases of $\widehat{V}(\widehat{\boldsymbol{\theta}})$ estimated with unequal covariance matrices when $n = 600$ are between those estimated with equal ones when $n = 150$ and 300; it indicates that a larger sample size is required to abate the nuisances due to over-parameterization. These tables further show that the estimates of asymptotic variances are fairly accurate. However, the bounds of normal-type confidence intervals seem to be easily overestimated, and these intervals are wider than the quantile ones. Again, the inference procedure based on the pooled sample covariance matrix outperforms that based on the group-specific sample covariance matrices. The results can be explained as a consequence of a high value of $V(\boldsymbol{\theta}_0)$ and a poor symmetric approximation to the left-skewed sampling distribution of $\widehat{V}(\widehat{\boldsymbol{\theta}})$ with small sample size.

4.3 Scenario III: Univariate Normal Marker

The last simulation was implemented to examine the rationale of sequential classification procedures. The univariate normal markers Y conditioning on $G = 1, 2$, and 3 were specified with means 0, 1, and 1.4, respectively, and the same standard deviation of 1. It is easy to justify that the designed models in the first setting satisfy the MLR condition. In addition, we consider markers with standard deviations $(1, 1.1, 1.3)$ as in the numerical study of Nakas and

Yiannoutsos (2004) and $(1, 2, 3)$, which correspond with a mild and a serious violation of the MLR. In this scenario, the HUM $V(\boldsymbol{\theta}_0)$ was estimated with sequential classification based on the original markers and optimal classification after transformation of likelihood ratios. Indeed, sequential classification achieves optimality only in the first setting. We note that the sequential procedure only requires to estimate the sample means in the classification rule

$$\widehat{G}(Y) = k_1 1(Y < c_1) + k_2 1(c_1 < Y < c_2) + k_3 1(Y > c_2)$$

with $c_1 < c_2$ and $\overline{Y}_{k_1} < \overline{Y}_{k_2} < \overline{Y}_{k_3}$. Without relying on the mechanism of MLR, the optimal classification needs to further estimate the variance of each class in the likelihood ratio transformation.

Tables 5 and 6 show that the influence of sample size on bias is similar to those in Sections 4.1 and 4.2. It is worth pointing out that the bounds of normal-type confidence intervals underestimate the true bounds, the tendency toward the side opposite to estimated bounds in the previous settings. This may similarly result from using the normal approximation to the right-skewed sampling distribution of $\widehat{V}(\widehat{\boldsymbol{\theta}})$ as $V(\boldsymbol{\theta}_0) < 0.5$ with small sample size. Misspecification of models for classification is also at issue in this scenario. With the validity of a MLR model, the sequential procedure takes advantage of estimating fewer parameters and gives more precise estimators, whereas the optimal procedure suffers from over-parameterization. In contrast, the absence of the MLR condition leads to a biased estimator of $V(\boldsymbol{\theta}_0)$, which loses the benefit from fewer parameter estimates and even has an unacceptable bias in estimation for $V(\boldsymbol{\theta}_0)$. Although marked variation is detected due to small sample sizes, performance of optimal classification procedure is free from the invalidity of a MLR model and the corresponding estimates actually move toward the true values. This simulation suggests that the validity of the MLR condition should be tested before conducting a sequential classification procedure.

5 Application to Hepatic Enzyme Profile

An analysis for liver function data was carried out to illustrate the usefulness of optimal ROC manifolds and HUMs. In most foregoing analyses, the data set of liver function tests in Albert and Harris (1987) was used to outline approaches of model fitting in discriminant analysis. Here we show that optimal ROC manifolds and HUMs will greatly assist in evaluating overall discriminability through outlining the capacity of enzymes for discrimination of viral hepatitis. Of these 218 patients, there are 57, 44, 40, and 77 ones diagnosed as acute viral hepatitis (Class 1), persistent chronic hepatitis (Class 2), aggressive chronic hepatitis (Class 3), and post-necrotic cirrhosis (Class 4), respectively. The four disease groups were determined by laparoscopy and biopsy. Four enzymes aspartate aminotrasferase (AST), alanine aminotransferase (ALT), glutamate dehydrogenase ($GLDH$), and ornithine carbonyltransferase (OCT) were collected and treated as biomarkers in classification. As Lesaffre and Albert (1989) suggested, these four variables were logarithmically transformed to remove skewness and outliers; a multinomial logistic regression model was adopted to fit $P(G = k|\{AST, ALT, GLDH\})$'s with the biomarker OCT being excluded in the following analysis due to its insignificant explanatory power.

For binary classification tasks, researchers have established some empirical rules of thumb for interpreting AUC, whereas no standard criterion can be followed in multi-classification. Using the relation between measures of a convex set and its projections, we suggest an estimated bounds of the HUM in K-classification based on performance of the marker in $K-1$ partial classifications. Since the set $\{\lambda\mathbf{p}_1 + (1-\lambda)\mathbf{p}_2 : \mathbf{p}_1 \in \phi_S(\mathcal{C}), \mathbf{p}_2 \in \mathcal{R}_S, p_{2,jk} = \delta_{jk}\}$ is convex and compact, some tools in convex geometry allow us to give a more precise range of V_S in terms of the HUMs in partial classifications. Based on a upper (UB) bound and a lower bound (LB) derived by Loomis and Whitney (1949) and Meyer (1988), respectively, with some modification, we have

$$\max\{\frac{1}{K!}, (\frac{K!}{K^K}\prod_{k=1}^{K} A_{S,k})^{1/(K-1)}\} \leq V_S \leq \min\{A_{S,1}, \ldots, A_{S,K}, \prod_{k=1}^{K} A_{S,k}^{1/(K-1)}\} \tag{16}$$

with $S = \{p_{11}, \ldots, p_{KK}\}$ and $A_{S,k}$ being the HUM in $S\backslash\{p_{kk}\}$-classification. The upper bound in (16) can be actually attained only when a marker is perfect in at least $K-2$ $(K-1)$-classifications, and the lower bound is attained only for performance of non-informative K-classification markers. These bounds will provide additional information to assess the performance of markers.

Table 7 and Figure 1 illustrate the total and partial discriminatory potential of the biomarker $(AST, ALT, GLDH)$. For each ternary classification, using the estimator in (2), the optimal ROC manifolds can be visualized in Figure 1 and convey sufficient information. The manifolds in Figure 1 (a) and (b) have shapes like unit cubes with the corresponding HUMs near 1. It suggests that classifiers based on the linear predictors have near perfect discrimination for patients in $\{1, 2, 3\}$ or $\{1, 2, 4\}$-classification. Comparatively, restricted by the ability of the biomarker in discerning aggressive chronic hepatitis and post-necrotic cirrhosis, the manifolds including the coordinates p_{33} and p_{44} take on cylinder-like shapes. Their HUM estimates are apparently reduced to 0.811 and 0.785. Even so, the performances are still closed to the theoretical upper bounds in (16). Their optimal ROC curves accompanied with confidence bands are included in the manifolds; see Figure 2 for instance. Nevertheless, for 4-classification, we are merely able to rely the accuracy summary to assess the classification capacity of the marker. Although each HUM in ternary classification is more than 0.8, V_S dramatically declines to 0.601. That is, even though it is actually well-performed in each 3-classification, the liver function test could not provide sufficient information and so the four disease groups have a overlap to some extent in the decision space. This example addresses that we could not capture the complete discriminability of a marker in quaternary-classification only through performances

in any partial ternary-classification.

Remark 5.1. When some false probabilities are concerned, a formula for V_S similar to (16) could not be obtained straightforward. For any nonempty set $\widetilde{S} \subset S = \{p_{11}, \ldots, p_{KK}\}$, performance of markers in $\#\widetilde{S}$-classification can be fully represented in $\mathcal{R}_{\widetilde{S}}$ and the relation between V_S and $V_{\widetilde{S}}$ will be established. Instead, admissibility of $\widetilde{S} \subset S = \{p_{k\sigma(k)}, k \neq \sigma(k)\}_{k=1}^{K}$ usually involves the number of classes more than the size of $\widetilde{S}$ and the performance set will degenerate to a space of a lower dimensionality; for instance, $\widetilde{S} = \{p_{12}, p_{23}\} \subset \{p_{12}, p_{23}, p_{31}\}$. Therefore, similar bounds either are too rough to be applied or do not exist at all.

6 Conclusive Discussion

We begin our discussion by reviewing the developed theoretical framework, and turn to different aspects of assessment of markers in multi-classification. This conclusion is devoted to stressing out new possibilities for future research.

6.1 Limitation of Prediction Probability

The performance probabilities p_{jk}'s play a pivotal role in the methodology we have established. Having this work, one may surmise that the prediction probabilities $P(G = k|\widehat{G} = j)$'s seem to be a plausible basis for an alternative evaluation of markers. After all, for binary classification, if $\widehat{G}_1 \succeq_S \widehat{G}_2$, it follows that

$$\frac{p_{11}(\widehat{G}_1)}{p_{11}(\widehat{G}_2)} \geq \frac{p_{11}(\widehat{G}_1)p_1 + p_{12}(\widehat{G}_1)p_2}{p_{11}(\widehat{G}_2)p_1 + p_{12}(\widehat{G}_2)p_2} \geq \frac{p_{12}(\widehat{G}_1)}{p_{12}(\widehat{G}_2)},$$

which implies that $P(G = 1|\widehat{G}_1 = 1) \geq P(G = 1|\widehat{G}_2 = 1)$ and $P(G = 2|\widehat{G}_1 = 1) \geq P(G = 2|\widehat{G}_2 = 1)$. Thus, an admissible classifier is an optimal predictor, which has the highest prediction probabilities, and vice versa. However, generally, the convexity of the set formed by prediction probabilities of all classifiers is valid only for $K = 2$, that makes characterization of

such a set complicated. Furthermore, estimates of $P(G = k|\widehat{G} = j)$'s are usually sensitive to noise as $P(\widehat{G} = j)$ moves toward zero. This will be inevitable to estimate prediction probabilities of those classifiers once one attempts to establish an overall evaluation of markers. Thus, even if we can surmount this theoretical impediment, the numerical instability in estimation for prediction probabilities often causes the lack of practicality in assessment.

6.2 Markers with Discrete or Mixture Distributions

Indeed, without assuming that $L_{kK}(\mathbf{Y})$'s are continuous random variables, the developed theory is still compatible with multi-classification tasks based on markers with discrete or mixture distributions. Once $P(L_{kK}(\mathbf{Y}) = c) > 0$ for some k and $c > 0$, it reveals that $L(\mathbf{Y})$ is a discrete or mixture random quantity. Consequently, there exists a subset of $\partial D_k(\mathbf{u})$ with positive probability mass in $\mathcal{D}$. Correspondingly, in ROC subspaces, $\phi_S(\mathcal{C})$ would lose its strict convexity, and some points in $M_S(\mathbf{u})$ form a flat part of $\partial\phi_S(\mathcal{C})$. To enhance the theoretical framework, we summarize its influence and make necessary modification. Specifically, it only needs a further investigation for $\widehat{G}_{\mathbf{u}}$ with $\phi_S(\widehat{G}_{\mathbf{u}})$ locating on the flat parts of M_S. First, the classifier $\widehat{G}_{\mathbf{u}}$ can be reformulated as

$$P(\widehat{G}_{\mathbf{u}}(\mathbf{Y}) = k|L(\mathbf{Y}) \in \widetilde{D}) = \begin{cases} 1 & \text{for } \widetilde{D} = D_k(\mathbf{u})\backslash\partial D_k(\mathbf{u}) \\ \lambda_{I,k} & \text{for } \widetilde{D} = \cap_{k\in I}\partial D_k(\mathbf{u}) \end{cases}$$

with $I \subset [K]$ and $\sum_{k\in I}\lambda_{I,k} = 1$. With continuous optimal markers, the admissibility induces a deterministic classification rule. Instead, an admissible classifier could be randomized since the connection between admissibility and utility criterion still maintains. Second, M_S remains to be a manifold under the considered types of markers. Since $M_S(\mathbf{u}) \subset H_S(\sup_{\widehat{G}\in\mathcal{C}}\mathbf{u}^\top\phi_S(\widehat{G}), \mathbf{u})$, the identity mapping from $M_S(\mathbf{u})$ to H_S ensures that the set $M_S(\mathbf{u})$ is structurally similar to Euclidean space, and its differentiable structure follows. For $\mathbf{Y}_m \in \cap_{k\in I}\partial D_k(\mathbf{u})$, $\widehat{G}_{\mathbf{u}}$ is randomly generated from a multinomial distribution with parameters $\lambda_{I,k}$'s. The empirical estimator of

M_S is naturally replaced by

$$\widehat{p}_{jk}(\widehat{G}_{\mathbf{u}}) = \widehat{n}_k^{-1} \sum_{m=1}^{n} 1(\widehat{G}_{\mathbf{u}}(\mathbf{Y}_m) = j)1(G_m = k).$$

Similar modification in the estimators can be made, and their functional asymptotic properties hold because the VC dimension of the family of $\widehat{G}_{\mathbf{u}}$ is unchanged. Finally, the characterization of HUM is still sound since the conditions we established before for the well-definition and well-behavior of HUM do not rely on the continuity assumption of $L(\mathbf{Y})$. Following the proof of Theorem 3.2, we can further verify the equality $HUM = CP$, although tediously, by separately considering the two sets of $L(\mathbf{Y})$ having probability mass or not. To avoid generating random quantity for $\widehat{G}(\mathbf{Y}_k)$'s, the kernel $h_S(\mathbf{Y}_1, \ldots, \mathbf{Y}_K)$ of $\widehat{V}$ can be slightly simplified as

$$\prod_{I \subset [K]} [1 + (\#I!^{-1} - 1)1(\exists\ \mathbf{u} : \cap_{k \notin I}\{L(\mathbf{Y}_k) \in D_{\sigma(k)}(\mathbf{u}) \backslash \partial D_{\sigma(k)}(\mathbf{u})\})$$
$$\cdot\, 1(\exists\ \mathbf{u} : L(\mathbf{Y}_k) \in \cap_{k \in I} \partial D_{\sigma(k)}(\mathbf{u}) \backslash \cup_{k \notin I} D_{\sigma(k)}(\mathbf{u}))].$$

Thus, the modified U-estimator for V is still consistent. However, for model-based inference, the asymptotic normality of $\widehat{V}(\widehat{\boldsymbol{\theta}})$ may not remain due to the violation of an applied model.

6.3 Comparisons among Optimal ROC Manifolds

Once optimal ROC manifolds are built up, we now return to the motivating interest in this article: drawing a fair comparison among markers. For this purpose, testing the equality between the HUMs V_1 of $\mathbf{Y}_1$ and V_2 of $\mathbf{Y}_2$ can be simply establish. Their optimal ROC manifolds M_1 and M_2 are different whenever $V_1 \neq V_2$. In addition, $V_1 = V_2$ ensures $M_1 = M_2$ if the two manifolds do not cross each other. Consider that $\mathbf{Y}_2$ contains only partial information of $\mathbf{Y}_1$; that is, $\mathbf{Y}_1 = g(\mathbf{Y}_2)$ for some function g. Since $\widehat{G}(\mathbf{Y}_1)$ can be rewritten as $\widehat{G}(g(\mathbf{Y}_2))$ for any $\widehat{G}$, there is no doubt about the relation $\phi_{\mathbf{Y}_1}(\mathcal{C}) \subset \phi_{\mathbf{Y}_2}(\mathcal{C})$. Thus, it suffices to evaluate the difference between V_1 and V_2. In application, researchers would be interested in comparison between $\mathbf{Y}_1$ and $\mathbf{Y}_2 = (\mathbf{Y}_1^\top, Y^*)^\top$. In this case, $V_2 - V_1$ can be interpreted as the marginal

discriminability of Y^*, which provides a reference for practitioners to decide the necessity of Y^* in classification. An estimator $\widehat{V}_2 - \widehat{V}_1$ for the difference $V_2 - V_1$ is naturally proposed and its asymptotic properties are inherited from $\widehat{V}_1$ and $\widehat{V}_2$.

Without the mentioned relation between $\mathbf{Y}_1$ and $\mathbf{Y}_2$, HUMs are unable to fully represent discrepancy between markers in performance. A justifiable approach becomes urgently necessary and important. It is naturally proposed to test $M^*_S(S\backslash p_{ij})$ via comparing p_{ij} with other performance probabilities being fixed, or $M_S(\mathbf{u})$ via examining the difference of performance of optimal classifiers over utility criteria $\mathbf{u}$. Similar to the inference for difference between HUMs, the discrepancy of the manifolds M_1 and M_2 can be estimated by $\widehat{M}^*_{S,1}(S\backslash p_{ij}) - \widehat{M}^*_{S,2}(S\backslash p_{ij})$ or $\widehat{M}_{S,1}(\mathbf{u}) - \widehat{M}_{S,2}(\mathbf{u})$. Their practicality and powers remain for future research.

References

Albert, A. and Harris, E. K. (1987). *Multivariate Interpretation of Clinical Laboratory Data*. CRC Press, first edition.

Beineke, L. W. and Wilson, R. J. (2004). *Topics in Algebraic Graph Theory*. Cambridge University Press, Cambridge.

Edwards, D. C., Metz, C. E., and Kupinski, M. A. (2004). Ideal observers and optimal ROC hypersurfaces in n-class classification. *IEEE Tans. on Med. Imag.*, 23(7):891–895.

Edwards, D. C., Metz, C. E., and Nishikawa, R. M. (2005). The hypervolume under the ROC hypersurface of "near-guessing" and "near-perfect" observers in n-class classification tasks. *IEEE Trans. on Med. Imag.*, 24(3):293–299.

He, X. and Frey, E. C. (2006). Three-class ROC analysis—the equal error utility assumption

and the optimality of three-class ROC surface using the ideal observer. *IEEE Trans. on Med. Imag.*, 25(8):979–986.

Jost, J. (2008). *Riemannian Geometry and Geometric Analysis*. Springer, New York, fifth edition.

Lesaffre, E. and Albert, A. (1989). Multiple-group logistic regression diagnostics. *J. Roy. Statist. Soc. Ser. C*, 38(3):425–440.

Lin, D. Y., Wei, L. J., Yang, I., and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *J. Roy. Statist. Soc. Ser. B*, 62(4):711–730.

Loomis, L. H. and Whitney, H. (1949). An inequality related to the isoperimetric inequality. *Bull. Amer. Math. Soc.*, 55(10):961–962.

Meyer, M. (1988). A volume inequality concerning sections of convex sets. *Bull. Lond. Math. Soc.*, 20(2):151 –155.

Nakas, C. T. and Yiannoutsos, C. T. (2004). Ordered multiple-class ROC analysis with continuous measurements. *Stat. Med.*, 23(22):3437–3449.

Randles, R. H. (1982). On the asymptotic normality of statistics with estimated parameters. *Ann. Statist.*, 10(2):462–474.

Schubert, C. M., Thorsen, S. N., and Oxley, M. E. (2011). The ROC manifold for classification systems. *Pattern Recognit.*, 44(2):350–362.

Scurfield, B. K. (1998). Generalization of the theory of signal detectability to n-event m-dimensional forced-choice tasks. *J. Math. Psych.*, 42(1):5–31.

Vaart, A. W. v. d. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

A Appendix

A.1 Proof of Lemma 3.1

Let $\mathbf{Y}^{(k)}$ be a marker from the kth class with its realization being denoted by $\mathbf{y}^{(k)}$. For the identity σ_0, suppose that $\{L(\mathbf{y}^{(k)})\}_{k=1}^K$ can be correctly classified by one classifier $\widehat{G}_{\mathbf{u}}$ with respective to $\mathbf{c}(\mathbf{u}) = (c_1, \ldots, c_{K-1})^\top$. That is, we have the inequalities

$$L_{ij}(\mathbf{y}^{(j)}) \leq c_i/c_j \leq L_{ij}(\mathbf{y}^{(i)}) \tag{17}$$

for all i, j with $i \leq K-1$ and $c_K = 1$. It follows from (17) that

$$L_{ij}(\mathbf{y}^{(j)}) \leq \prod_{k=0}^{o(\sigma)-1} \frac{c_{\sigma^k(i)}}{c_{\sigma^{k+1}(i)}} \leq L_{ij}(\mathbf{y}^{(i)}) \tag{18}$$

for non-identity σ's, where $o(\sigma) = \arg\min_k\{\sigma^k(j) = j \ \forall j\}$. Trough tedious algebra, the roles of c_i's in (18) can be eliminated and the inequalities can be further simplified as

$$\prod_{k=0}^{o(\sigma)-1} L_{\sigma^k(i)\sigma^{k+1}(i)}(\mathbf{y}^{(\sigma^k(i))}) \geq 1,$$

which implies that $\prod_{k=1}^K f_k(\mathbf{y}^{(k)}) \geq \prod_{k=1}^K f_k(\mathbf{y}^{(\sigma(k))})$. Similarly, the proof for the converse part can be completed by showing

$$\bigcap_{\sigma}\{\mathbf{c} : c_i/c_{\sigma(i)} \in [L_{\sigma(i)i}(\mathbf{y}^{(\sigma(i))}), \prod_{k=0}^{o(\sigma)-1} L_{\sigma^k(i)\sigma^{k+1}(i)}(\mathbf{y}^{(i)})]\} \neq \emptyset.$$

As for the general permutation σ_0, we first substitute $\mathbf{y}^{(k)}$ by $\mathbf{y}^{(\sigma_0(k))}$ and σ by $\sigma(\sigma_0^{-1})$. Along the same lines as the above proof, the lemma is directly obtained.

A.2 Proof of Theorem 3.2

Without loss of generality, we consider $S = \{p_{11}, \ldots, p_{KK}\}$. Let $\boldsymbol{\ell} = L(\mathbf{y})$ and $\boldsymbol{\ell}^{(k)} = L(\mathbf{y}^{(k)})$. For $S = \{p_{k\sigma(k)} : \sigma(k) \neq k\}_{k=1}^K$, the proof can be accomplished in the same manner by replacing $\mathbf{y}^{(k)}$ with $\mathbf{y}^{(\sigma(k))}$. By definition, the HUM can be expressed as an integral with respect to the

critical point $\mathbf{c}(\mathbf{u}) \in \mathcal{D}$ as follows:

$$V_S = \int_{\mathbb{R}^{K-1}} p_{KK}(\widehat{G}_{\mathbf{u}}(\mathbf{c}))|\det \mathbf{J}(\mathbf{c})|d\mathbf{c}, \tag{19}$$

where $\mathbf{J}(\mathbf{c}) = \partial_{\mathbf{c}}(p_{11}(\mathbf{c}), \ldots, p_{(K-1)(K-1)}(\mathbf{c}))$. Here, the (i,j)th element $J_{ij}(\mathbf{c})$ in $\mathbf{J}(\mathbf{c})$ can be explicitly derived to be

$$J_{ij}(\mathbf{c}) = \int_{H_{ij}(\mathbf{c})} f_{L_i}(\boldsymbol{\ell})d\boldsymbol{\ell} - \delta_{ij}\sum_{k=1}^{K}\int_{H_{ik}(\mathbf{c})} f_{L_i}(\boldsymbol{\ell})d\boldsymbol{\ell}$$

with $H_{ij}(\mathbf{c}) = D_i(\mathbf{u}(\mathbf{c})) \cap D_j(\mathbf{u}(\mathbf{c}))$. Since it is a weighted Laplacian submatrix in the sense of graph theory, the matrix-tree theorem (Beineke and Wilson, 2004) enables us to calculate its determinant as

$$\begin{aligned}\det \mathbf{J}(\mathbf{c}) &= (-1)^K \sum_{T\subset G[K]} \prod_{(i,j)\in T} \int_{H_{ij}(\mathbf{c})} f_{L_i}(\boldsymbol{\ell}^{(i)})d\boldsymbol{\ell}^{(i)} \\ &= (-1)^K \int_{H(\mathbf{c})} \prod_{k=1}^{K-1} f_{L_k}(\boldsymbol{\ell}^{(k)})d(\boldsymbol{\ell}^{(1)}, \ldots, \boldsymbol{\ell}^{(K-1)}),\end{aligned} \tag{20}$$

where T is a spanning tree of $G[K]$, a directed graph with the vertex set $[K]$ and the weighted edges $J_{ij}(\mathbf{c})$, and $H(\mathbf{c}) = \bigcup_{T\in G[K]} \times_{(i,j)\in T} H_{ij}(\mathbf{c})$. By replacing $\det \mathbf{J}(\mathbf{c})$ in (19) with (20), we can immediately obtain that

$$V_S = \int_{H(\mathbf{c})\times D_K(\mathbf{u}(\mathbf{c}))} f_{\mathbf{X}}(\mathbf{x})d\mathbf{x},$$

where $\mathbf{X} = (L(\mathbf{Y}^{(1)}), \ldots, L(\mathbf{Y}^{(K)}))^\top$ and $f_{\mathbf{X}}$ denotes the joint probability density function of $\mathbf{X}$. Moreover, it follows from Lemma 3.1 that the remaining work of this proof is to demonstrate the equality

$$\{(\mathbf{c}, \mathbf{x}) : \mathbf{x} \in H(\mathbf{c}) \times D_K(\mathbf{u}(\mathbf{c}))\} = \{(\mathbf{c}, \mathbf{x}) : \boldsymbol{\ell}^{(k)} \in D_k(\mathbf{u}(\mathbf{c})) \ \forall k = 1, \ldots, K\}.$$

Obviously, the containing relation "$\subset$" is trivial. Conversely, for given realization $\mathbf{x}$, we specify

$$\mathbf{c} = \arg\max_{\mathbf{c}^*}\{P(L(\mathbf{Y}^{(K)}) \in D_K(\mathbf{u}(\mathbf{c}^*))) : c_i^*/c_j^* \in [\ell_i^{(j)}/\ell_j^{(j)}, \ell_i^{(i)}/\ell_j^{(i)}]^\top \forall i, j, c_K^* = 1\}.$$

Let G^* be a graph with the vertex set $[K]$ and the edge set $\{(i,j) : \boldsymbol{\ell}^{(i)} \in H_{ij}(\mathbf{c}), i = 1, \ldots, K-1\}$. Suppose that G^* contains no tree as its subgraph and so some vertex, say i_0, is isolated.

There exists $\mathbf{c}_\delta = (c_1, \ldots, c_{i_0} + \delta, \ldots, c_K)^\top$ with $\delta > 0$ satisfying $c_{\delta i}/c_{\delta j} \in [\ell_i^{(j)}/\ell_j^{(j)}, \ell_i^{(i)}/\ell_j^{(i)}]$ for all i, j, and $P(L(\mathbf{Y}^{(K)}) \in D_K(\mathbf{u}(\mathbf{c}_\delta))) > P(L(\mathbf{Y}^{(K)}) \in D_K(\mathbf{u}(\mathbf{c}^*)))$. This contradicts what we suppose and, hence, $\mathbf{x} \in H(\mathbf{c}) \times D_K(\mathbf{u}(\mathbf{c}))$. Thus, the proof is completed.

A.3 Proof of Theorem 3.3

By the law of large numbers and the compactness of $\boldsymbol{\Theta}$, it yields that

$$\widehat{V}(\boldsymbol{\theta}) = \widehat{V}^*(\boldsymbol{\theta})(1 + r_n(\boldsymbol{\theta})), \tag{21}$$

where

$$\widehat{V}^*(\boldsymbol{\theta}) = (\prod_{k=1}^{K} n_k)^{-1} \sum_{\{i_1,\ldots,i_K\}\subset[n]} h_S(\mathbf{Y}_{i_1}, \ldots, \mathbf{Y}_{i_K}; \boldsymbol{\theta}) \prod_{k=1}^{K} 1(G_{i_k} = k)$$

and $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |r_n(\boldsymbol{\theta})| = o_p(1)$. With the asymptotic equivalence in (21), the remaining work is to establish the asymptotic normality of $n^{1/2}(\widehat{V}^*(\widehat{\boldsymbol{\theta}}) - \widehat{V}^*(\boldsymbol{\theta}_0))$. Define

$$Q(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\theta}') = |h_S(\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(K)}; \boldsymbol{\theta}) - h_S(\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(K)}; \boldsymbol{\theta}')|.$$

By Theorem 2.13 in Randles (1982), it suffices to show that for any $\varepsilon > 0$

$$\mathrm{E}[\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|<\varepsilon} Q(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\theta}')] \le \lambda_0 \varepsilon \tag{22}$$

and

$$\lim_{\varepsilon\to 0^+} \mathrm{E}[\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|<\varepsilon} Q^2(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\theta}')^2] = 0 \tag{23}$$

for some constant λ_0. Obviously, (23) is a direct result of (22) because $Q(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\theta}')$ is bounded.

Since a sufficient condition for $Q(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\theta}') = 0$ can be stated as that there exists $\mathbf{c}' = (c'_1, \ldots, c'_{K-1})^\top$ such that both $g(\mathbf{y}^{(k)}; \boldsymbol{\theta})$ and $g(\mathbf{y}^{(k)}; \boldsymbol{\theta}')$ belong to $D_{\sigma(k)}(\mathbf{u}(\mathbf{c}))$ for some σ and all $\mathbf{c}$ satisfying $e^{-\lambda_1\varepsilon} < (\sum_{k=1}^{K-1} c_k^2/c_k'^2)^{1/2} < e^{\lambda_1\varepsilon}$ for some $\lambda_1 > 0$. Similar to (17), such a $\mathbf{c}'$ exists if

$$g_{jk}(\mathbf{y}^{(k)}; \boldsymbol{\theta}) e^{-\lambda_1\varepsilon} \le c_j/c_k \le g_{jk}(\mathbf{y}^{(j)}; \boldsymbol{\theta}) e^{\lambda_1\varepsilon}. \tag{24}$$

Along the same lines as the proof of Lemma 3.1, the inequality in (24) can be ensured through the simplified condition

$$\prod_{k=0}^{o(\sigma)-1} g_{\sigma^k(i)\sigma^{k+1}(i)}(\mathbf{y}^{(\sigma^k(i))};\boldsymbol{\theta}) \geq e^{2K\lambda_1\varepsilon} \tag{25}$$

for all non-identity σ's. Hence, $Q(\mathbf{X},\boldsymbol{\theta},\boldsymbol{\theta}')=1$ implies the negation of (25) as

$$e^{-2K\lambda_1\varepsilon} < \prod_{k=1}^{K} g_{k\sigma(k)}(\mathbf{y}^{(k)};\boldsymbol{\theta}) < e^{2K\lambda_1\varepsilon} \tag{26}$$

for all σ. Let $\mathcal{D}_M = \mathcal{D}\cap B_M(0)\backslash B_{1/M}(0)$ satisfying $P(g(\mathbf{Y}^{(k)};\boldsymbol{\theta})\notin\mathcal{D}_M)<\eta$ for given $\eta>0$, where $B_M(0)$ is a ball centered at the origin and with radius $M>1$. We can further derive from (26) that

$$\begin{aligned}
&\mathrm{E}[Q(\mathbf{X},\boldsymbol{\theta},\boldsymbol{\theta}')] \qquad (27)\\
&\leq \mathrm{E}[Q(\mathbf{X},\boldsymbol{\theta},\boldsymbol{\theta}')\prod_{k=1}^{K} 1(g(\mathbf{Y}^{(k)};\boldsymbol{\theta})\in\mathcal{D}_M)] + P(\exists\, g(\mathbf{Y}^{(k)};\boldsymbol{\theta})\notin\mathcal{D}_M)\\
&\leq P(g(\mathbf{Y}^{(k)};\boldsymbol{\theta})\in\cap_{k=1}^{K-1}H_{k,\varepsilon}(\mathbf{c})\cap\mathcal{D}_M) + K\eta\\
&< (2M)^{K-2}(K-1)P(\|\ln\frac{g(\mathbf{Y}^{(k)};\boldsymbol{\theta})}{g(\mathbf{Y}^{(k)};\boldsymbol{\theta}')}\|<4K\lambda_1\varepsilon) + K\eta,
\end{aligned}$$

where $H_{k,\varepsilon}(\mathbf{c}) = \{\mathbf{c}' : |c_k/c_k'| < e^{2K\lambda_1\varepsilon}\}$ and $\boldsymbol{c}$ can be determined by $\{\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(K-1)}\}$. By setting $\lambda_0 = 2^K M^{K-2}K(K-1)\lambda_1$ in (27) and choosing a suitable λ_1 for a log-Lipschitz continuoity, the inequality (22) is automatically satisfied. The compactness of $\boldsymbol{\Theta}$ and $\inf_{\mathbf{y}} g(\mathbf{y};\boldsymbol{\theta})>0$ on $\mathcal{D}_M$ ensure the log-Lipschitz continuity from the Lipschitz continuity and, hence, one can conclude that

$$\sqrt{n}(\widehat{V}^*(\widehat{\boldsymbol{\theta}}) - V^*(\boldsymbol{\theta}_0)) \xrightarrow{d} N(0,\xi(\boldsymbol{\theta}_0)).$$

Together with (21), the asymptotic normality of $\widehat{V}(\widehat{\boldsymbol{\theta}})$ is as claimed.

Table 1: Estimation for $V(\boldsymbol{\theta}_0)$ by $\widehat{V}(\boldsymbol{\theta}_0)$ under multinomial logistic regression

$(\rho, V(\boldsymbol{\theta}_0))$	$(0.2, 0.672)$			$(0.5, 0.616)$			$(0.8, 0.517)$		
n	$mean$	sd	se	$mean$	sd	se	$mean$	sd	se
150	0.671	0.0527	0.0512	0.618	0.0517	0.0526	0.516	0.0534	0.0535
300	0.671	0.0359	0.0360	0.615	0.0373	0.0371	0.517	0.0374	0.0377
450	0.672	0.0291	0.0293	0.616	0.0302	0.0302	0.517	0.0315	0.0307
600	0.672	0.0256	0.0253	0.616	0.0259	0.0262	0.517	0.0264	0.0266

Table 2: Estimation and inference procedures for $V(\boldsymbol{\theta}_0)$ by $\widehat{V}(\widehat{\boldsymbol{\theta}})$ under multinomial logistic regression

ρ	n	$mean$	sd	se	QI	L_q	CI_z	L_z	E_{cp}
0.2	150	0.684	0.0521	0.0503	(0.5830,0.7839)	0.2009	(0.5820,0.7862)	0.2042	0.924
	300	0.678	0.0356	0.0357	(0.6047,0.7472)	0.1426	(0.6084,0.7481)	0.1397	0.940
	450	0.676	0.0289	0.0292	(0.6203,0.7318)	0.1115	(0.6198,0.7330)	0.1132	0.944
	600	0.676	0.0254	0.0253	(0.6253,0.7252)	0.0999	(0.6259,0.7256)	0.0997	0.939
0.5	150	0.633	0.0514	0.0520	(0.5299,0.7356)	0.2057	(0.5319,0.7334)	0.2014	0.928
	300	0.623	0.0370	0.0369	(0.5482,0.6935)	0.1453	(0.5501,0.6953)	0.1452	0.940
	450	0.621	0.0302	0.0301	(0.5626,0.6792)	0.1166	(0.5619,0.6802)	0.1183	0.937
	600	0.620	0.0258	0.0261	(0.5689,0.6693)	0.1003	(0.5691,0.6704)	0.1013	0.948
0.8	150	0.533	0.0527	0.0533	(0.4311,0.6369)	0.2057	(0.4299,0.6364)	0.2065	0.935
	300	0.525	0.0373	0.0376	(0.4516,0.5987)	0.1471	(0.4519,0.5983)	0.1464	0.943
	450	0.522	0.0313	0.0307	(0.4638,0.5857)	0.1219	(0.4610,0.5837)	0.1227	0.942
	600	0.521	0.0262	0.0266	(0.4701,0.5707)	0.1006	(0.4694,0.5722)	0.1028	0.956

Table 3: Estimation for $V(\boldsymbol{\theta}_0)$ by $\widehat{V}(\boldsymbol{\theta}_0)$ under the multivariate normal markers with the common covariance matrix

$(\rho, V(\boldsymbol{\theta}_0))$	$(0.2, 0.614)$			$(0.5, 0.664)$			$(0.8, 0.831)$		
n	$mean$	sd	se	$mean$	sd	se	$mean$	sd	se
150	0.615	0.0530	0.0523	0.665	0.0507	0.0499	0.833	0.0360	0.0362
300	0.614	0.0365	0.0369	0.665	0.0350	0.0352	0.833	0.0254	0.0255
450	0.614	0.0304	0.0301	0.665	0.0285	0.0287	0.832	0.0208	0.0208
600	0.614	0.0263	0.0260	0.664	0.0245	0.0248	0.830	0.0182	0.0181

Table 4: Estimation and inference procedures for $V(\boldsymbol{\theta}_0)$ by $\widehat{V}(\widehat{\boldsymbol{\theta}})$ with pooled and group-specific covariance estimates under the multivariate normal markers with the common covariance matrix

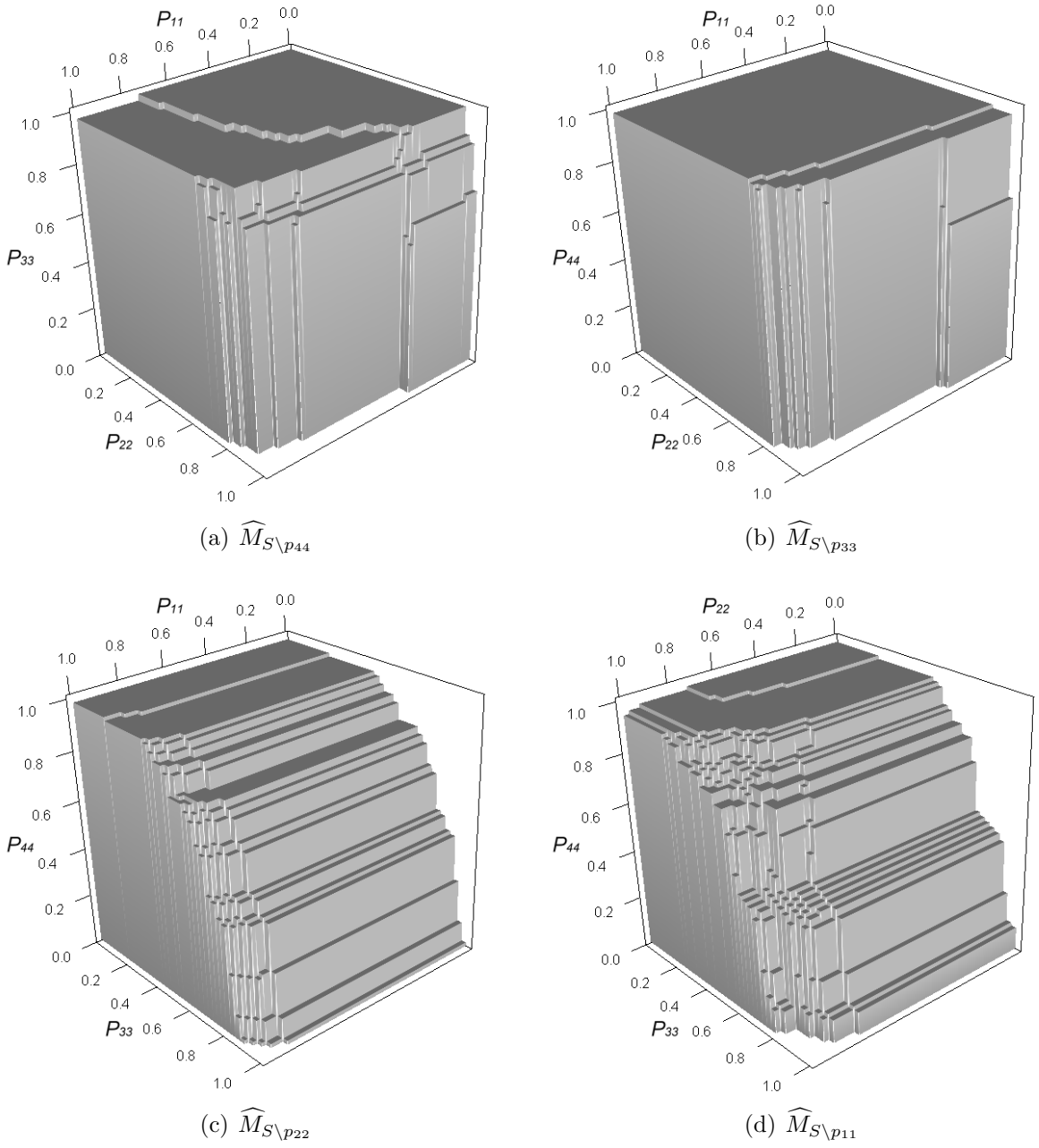
		Pooled Covariance Matrix Estimate							
ρ	n	$mean$	sd	se	QI	L_q	CI_z	L_z	E_{cp}
0.2	150	0.626	0.0525	0.0517	(0.5237,0.7262)	0.2025	(0.5227,0.7286)	0.2060	0.930
	300	0.619	0.0365	0.0367	(0.5486,0.6904)	0.1418	(0.5477,0.6908)	0.1431	0.950
	450	0.618	0.0303	0.0300	(0.5586,0.6756)	0.1170	(0.5586,0.6774)	0.1188	0.945
	600	0.617	0.0261	0.0260	(0.5652,0.6679)	0.1027	(0.5658,0.6682)	0.1023	0.946
0.5	150	0.674	0.0502	0.0494	(0.5754,0.7659)	0.1906	(0.5758,0.7725)	0.1967	0.936
	300	0.670	0.0350	0.0350	(0.6006,0.7368)	0.1362	(0.6015,0.7386)	0.1371	0.946
	450	0.668	0.0284	0.0286	(0.6117,0.7235)	0.1117	(0.6127,0.7240)	0.1114	0.944
	600	0.667	0.0245	0.0248	(0.6185,0.7137)	0.0952	(0.6187,0.7149)	0.0962	0.952
0.8	150	0.837	0.0357	0.0358	(0.7630,0.9012)	0.1382	(0.7669,0.9068)	0.1399	0.920
	300	0.833	0.0254	0.0255	(0.7816,0.8809)	0.0994	(0.7829,0.8826)	0.0998	0.937
	450	0.833	0.0207	0.0207	(0.7920,0.8729)	0.0809	(0.7924,0.8735)	0.0811	0.945
	600	0.831	0.0181	0.0180	(0.7953,0.8652)	0.0699	(0.7956,0.8664)	0.0709	0.947
		Group-Specific Covariance Matrix Estimate							
ρ	n	$mean$	sd	se	QI	L_q	CI_z	L_z	E_{cp}
0.2	150	0.647	0.0512	0.0507	(0.5474,0.7473)	0.1999	(0.5463,0.7468)	0.2005	0.877
	300	0.629	0.0360	0.0364	(0.5588,0.6983)	0.1394	(0.5586,0.6999)	0.1413	0.928
	450	0.625	0.0300	0.0298	(0.5649,0.6815)	0.1167	(0.5659,0.6835)	0.1175	0.933
	600	0.622	0.0260	0.0258	(0.5702,0.6722)	0.1020	(0.5711,0.6729)	0.1017	0.935
0.5	150	0.694	0.0483	0.0481	(0.5988,0.7852)	0.1864	(0.5997,0.7890)	0.1893	0.878
	300	0.680	0.0344	0.0346	(0.6125,0.7439)	0.1314	(0.6120,0.7470)	0.1350	0.930
	450	0.674	0.0281	0.0284	(0.6196,0.7298)	0.1102	(0.6193,0.7296)	0.1102	0.932
	600	0.672	0.0245	0.0246	(0.6237,0.7180)	0.0944	(0.6235,0.7195)	0.0960	0.940
0.8	150	0.849	0.0340	0.0341	(0.7748,0.9108)	0.1360	(0.7824,0.9159)	0.1335	0.880
	300	0.839	0.0250	0.0249	(0.7887,0.8852)	0.0965	(0.7899,0.8878)	0.0978	0.914
	450	0.837	0.0205	0.0204	(0.7967,0.8761)	0.0794	(0.7968,0.8770)	0.0802	0.918
	600	0.834	0.0179	0.0178	(0.7986,0.8678)	0.0692	(0.7989,0.8691)	0.0702	0.937

Table 5: Estimation for $V(\boldsymbol{\theta}_0)$ by $\widehat{V}(\boldsymbol{\theta}_0)$ based on optimal classification procedures under a MLR model and non-MLR models

MLR		True			Mild Violation			Serious Violation		
$V(\boldsymbol{\theta}_0)$		0.408			0.387			0.406		
n		*mean*	*sd*	*se*	*mean*	*sd*	*se*	*mean*	*sd*	*se*
150		0.410	0.0482	0.0481	0.386	0.0484	0.0482	0.408	0.0502	0.0505
300		0.407	0.0337	0.0337	0.388	0.0340	0.0338	0.406	0.0354	0.0354
450		0.407	0.0270	0.0274	0.388	0.0277	0.0276	0.405	0.0289	0.0289
600		0.407	0.0234	0.0237	0.387	0.0242	0.0238	0.407	0.0254	0.0250

Table 6: Estimation and inference procedures for $V(\boldsymbol{\theta}_0)$ by $\widehat{V}(\widehat{\boldsymbol{\theta}})$ based on optimal and sequential classification procedures under a MLR model and non-MLR models

		Optimal Classification							
MLR	n	*mean*	*sd*	*se*	QI	L_q	CI_z	L_z	E_{cp}
True	150	0.414	0.0472	0.0485	(0.3253,0.5111)	0.1858	(0.3210,0.5061)	0.1851	0.955
	300	0.408	0.0336	0.0338	(0.3446,0.4758)	0.1312	(0.3422,0.4740)	0.1318	0.950
	450	0.408	0.0276	0.0274	(0.3542,0.4615)	0.1073	(0.3542,0.4623)	0.1081	0.950
	600	0.407	0.0234	0.0237	(0.3623,0.4528)	0.0905	(0.3614,0.4529)	0.0916	0.954
Mild Vio.	150	0.392	0.0480	0.0485	(0.3017,0.4877)	0.1859	(0.2976,0.4856)	0.1880	0.950
	300	0.390	0.0342	0.0340	(0.3251,0.4580)	0.1328	(0.3230,0.4570)	0.1340	0.945
	450	0.389	0.0277	0.0276	(0.3374,0.4450)	0.1076	(0.3347,0.4432)	0.1086	0.950
	600	0.388	0.0244	0.0239	(0.3425,0.4379)	0.0954	(0.3400,0.4356)	0.0956	0.948
Serious Vio.	150	0.417	0.0493	0.0508	(0.3235,0.5129)	0.1893	(0.3200,0.5133)	0.1933	0.956
	300	0.410	0.0349	0.0354	(0.3432,0.4790)	0.1358	(0.3417,0.4787)	0.1369	0.952
	450	0.408	0.0288	0.0289	(0.3528,0.4654)	0.1125	(0.3514,0.4645)	0.1131	0.949
	600	0.409	0.0253	0.0250	(0.3627,0.4587)	0.0960	(0.3590,0.4583)	0.0993	0.950
		Sequential Classification							
MLR	n	*mean*	*sd*	*se*	QI	L_q	CI_z	L_z	E_{cp}
True	150	0.410	0.0478	0.0481	(0.3206,0.5055)	0.1849	(0.3164,0.5038)	0.1874	0.949
	300	0.407	0.0337	0.0337	(0.3446,0.4758)	0.1312	(0.3412,0.4732)	0.1321	0.948
	450	0.408	0.0275	0.0274	(0.3546,0.4615)	0.1068	(0.3540,0.4619)	0.1079	0.950
	600	0.407	0.0234	0.0237	(0.3628,0.4527)	0.0899	(0.3612,0.4527)	0.0915	0.954
Mild Vio.	150	0.386	0.0487	0.0476	(0.2903,0.4835)	0.1932	(0.2903,0.4811)	0.1908	0.937
	300	0.387	0.0341	0.0335	(0.3219,0.4569)	0.1350	(0.3201,0.4539)	0.1338	0.940
	450	0.387	0.0276	0.0273	(0.3355,0.4430)	0.1075	(0.3333,0.4416)	0.1083	0.948
	600	0.386	0.0240	0.0236	(0.3419,0.4346)	0.0927	(0.3394,0.4334)	0.0939	0.947
Serious Vio.	150	0.283	0.0600	0.0438	(0.1580,0.3891)	0.2311	(0.1655,0.4007)	0.2352	0.310
	300	0.284	0.0470	0.0309	(0.1683,0.3565)	0.1882	(0.1919,0.3761)	0.1842	0.068
	450	0.287	0.0382	0.0255	(0.1812,0.3462)	0.1650	(0.2126,0.3622)	0.1496	0.017
	600	0.290	0.0330	0.0221	(0.1863,0.3390)	0.1528	(0.2248,0.3542)	0.1294	0.004



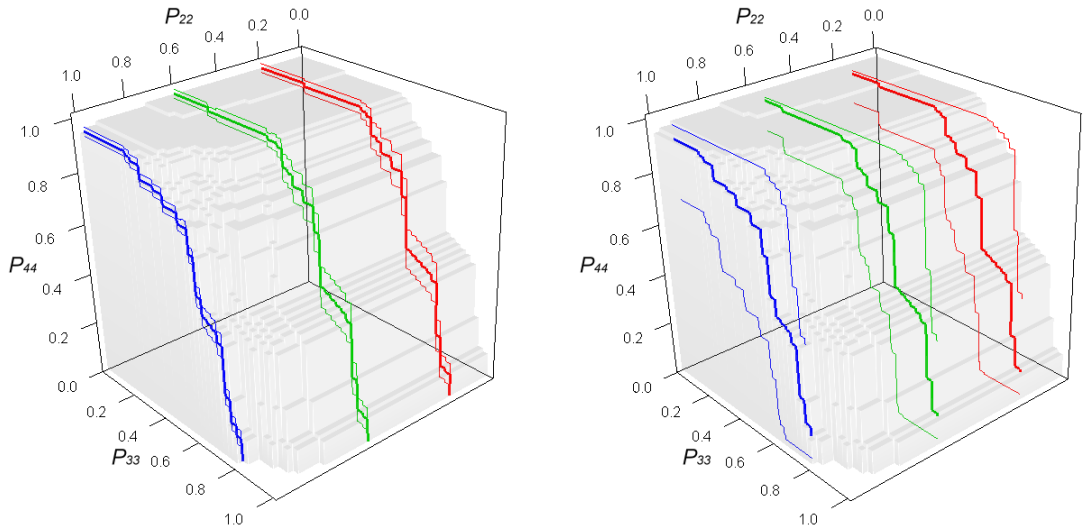
(a) $\widehat{M}_{S\setminus p_{44}}$ (b) $\widehat{M}_{S\setminus p_{33}}$

(c) $\widehat{M}_{S\setminus p_{22}}$ (d) $\widehat{M}_{S\setminus p_{11}}$

Figure 1: Estimated optimal ROC manifolds based on AST, ALT, and $GLDH$

Table 7: Estimates for V_S based on optimal transformations of AST, ALT, and $GLDH$

Classes	$\widehat{V}_S$	se	LB	UB	Classes	$\widehat{V}_S$	se
{1,2,3,4}	0.601	0.0520	0.376	0.785	{1,2}	0.961	0.0217
{1,2,3}	0.942	0.0254	0.457	0.961	{1,3}	0.995	0.0043
{1,2,4}	0.944	0.0264	0.455	0.961	{1,4}	0.987	0.0129
{1,3,4}	0.811	0.0446	0.421	0.814	{2,3}	0.981	0.0103
{2,3,4}	0.785	0.0457	0.418	0.814	{2,4}	0.983	0.0132
					{3,4}	0.814	0.0449



(a) Pointwise confidence bands (3) (b) Simultaneous confidence bands (7)

Figure 2: Estimated optimal ROC curves (solid line) for p_{33} and p_{44} with $p_{22} = 0.20, 0.50$, and 0.98 embedded in $\widehat{M_S}$ and the 0.95 confidence bands of $M_S \cap \{p_{kk} \geq 0.1, k = 2, 3, 4\}$ around the curves