Pattern Recognition 44 (2011) 350–362

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

The ROC manifold for classification systems[☆]

Christine M. Schubert[a,*], Steven N. Thorsen[b], Mark E. Oxley[a]

[a] Department of Mathematics and Statistics, Graduate School of Engineering and Management, Air Force Institute of Technology, 2950 Hobson Way, Wright-Patterson AFB, OH 45433-7765, United States
[b] Department of Mathematical Sciences, United States Air Force Academy, 2354 Fairchild Drive, Colorado Springs, CO 80840, United States

ARTICLE INFO

Article history:
Received 21 November 2008
Received in revised form
28 May 2010
Accepted 29 July 2010

Keywords:
Classification
Multiple classes
Receiver operating characteristic (ROC) curve
ROC manifold
Bayes cost

ABSTRACT

We define the ROC manifold and CC manifold as duals in a given sense. Their analysis is required to describe the classification system. We propose a mathematical definition based on vector space methods to describe both. The ROC manifolds for n-class classification systems fully describe each system in terms of its misclassifications and, by conjunction, its correct classifications. Optimal points which minimize misclassifications can be identified even when costs and prior probabilities differ. These manifolds can be used to determine the usefulness of a classification system based on a given performance criterion. Many performance functionals (such as summary statistics) preferred for CC manifolds can also be evaluated using the ROC manifold (under certain constraints). Examples using the ROC manifold and performance functionals to compete classification systems are demonstrated with simulated and applied disease detection data.

Published by Elsevier Ltd.

1. Introduction

Paramount to the development of classification systems is the ability to judge the usefulness of the system, whether judging the system against a benchmark level of acceptable performance or comparing it to other candidate systems. To make this judgement a performance criterion is required. One of the oldest and most commonly used performance tools used in the analysis of classification systems is the receiver operator characteristic (ROC) curve. The most commonly used ROC curve depicts the trade-off in correct classification for one pivotal class with the false classification into that class. A less common ROC curve depicts the trade-off of the two types of false classifications that can occur.

In the last decade, complexity in classification applications has warranted an extension of ROC curves and their analyses to describe and analyze systems in which there are three or more classes [1–11,14]. These extensions of ROC curves have produced various surfaces defined in terms of the correct classifications with the notable exception of [8,9,14], in which surfaces related to the misclassification errors are described. Points lying on these surfaces correspond to different operating parameters associated with the classification system. Often these parameters are thresholds (one example would be signal-to-noise ratio), though they need not be. There is no standardization of these surfaces and most focus on permutations of the correct classifications. For classification systems with three classes, these surfaces may be visualized in a three-dimensional plot of the true (correct) classification rates [1–7]. Since these surfaces are topological manifolds, we refer to them as correct classification manifolds (CC manifolds). For $n>2$ classes, concepts related to these surfaces have been proposed, many still focusing on the correct classification rates, though the increased dimensionality makes it impossible to view all correct classifications simultaneously [10]. At best for the n-class system, sets of three-dimensional plots can be used to examine the correct classifications for three classes at a time.

Initially, focusing on correct classification rates seems appealing since, for the three-class classification system, the trade-off between correct classifications can be compared graphically using each class's correct classification as an axis. Furthermore, summary measures of these CC manifolds focus on how well the classification systems correctly classify into their class states, thereby describing the overall correct classification rate. By conjunction, then, the overall misclassification rate for the entire system is described, although no information is directly obtainable about misclassifications within each class. Such summary measures include the total correct classification rate and volume under the surface (VUS) [1,6,9]. Many researchers have examined VUS for systems with more than two classes [12,3,10,13,6,14,4] with the view of constructing a polytope from the data to

[☆] The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the US Government.

[*] Corresponding author.

E-mail addresses: christine.schubertkabban@afit.edu (C.M. Schubert), Steven.Thorsen@usafa.edu (S.N. Thorsen), Mark.Oxley@afit.edu (M.E. Oxley).

0031-3203/$ - see front matter Published by Elsevier Ltd.
doi:10.1016/j.patcog.2010.07.025

calculate or describe how to use the VUS. The appeal of VUS is that this summary performance quantifier hopefully becomes a probability estimate as it does with the two-class case, generalizing the diagnostic ability of the classification system across all operating parameters. For a CC manifold, this can be interpreted as the chance of correct classification when presented with, as a group, one randomly selected subject from each class [1,6]. To illustrate further using medical diagnostics, the resulting VUS for three classes may be interpreted as the probability that a clinician diagnoses each individual to the correct diagnostic class after being presented with three individuals from three different classes. Herein lies more confusion, however, because in the VUS defined by the volume under a polytope created in the space based on classification errors, the probability of correct classification is not necessarily 1-VUS, if VUS indeed exists. In contrast, for the VUS defined by the volume under a polytope created in the space based on correct classifications, the probability of correct classification is the VUS.

There are noted issues surrounding the use of VUS [15,9]. In [15] we see that conclusions made when comparing classification systems based on VUS infer the classification system's diagnostic ability, with the caveat that these calculations assume equal weighting for prior probabilities and costs between the classes. However, there are no costs associated with correct classifications, only errors in classifications, and as such, summary statistics not considering misclassifications cannot address these costs. In [8,9] we see a definition of a ROC hypersurface and the hypervolume under it which extended previous efforts beyond the three-class case to an n-class case. It is demonstrated that the "guessing" (and through convergence the "near-guessing") observer has the same VUS as the "perfect" ideal observer.

As a result of these works, there are two important issues to address. First, there is a dual problem in the CC manifold. Given an n-class classification system, analysis of the dual problem involves an $(n-1)$-dimensional linear variety of the n-space containing the CC manifold. Since this linear variety is codimension 1 to the correct classification space (CC space), a surface can always be generated under it (ignoring the second issue discussed below). Therefore, [9] would have the ROC hypersurface VUS of every "perfect" ideal and "guessing" observer equal to 0. However, the CC manifold VUS of the "perfect" ideal observer is 1 in every case. This occurs because the surface created by the "guessing" observer will always be an n-simplex for this observer. For example, in the simpler two-class system which produces a ROC curve, we have $n=2$. Hence, the ROC space is in a space of dimension $2(2-1)=2$ while the ROC curve is isomorphic to a subset of the space $\mathbb{R}^1$, a space of dimension $2-1=1$, making the curve codimension 1 to the original ROC space. This creates a "volume" under the ROC curve. Notice also that the CC curve for the two-class system is also isomorphic to $\mathbb{R}^1$, which is codimension 1 to CC space. Thus, it too has a "volume". Of course, in these dimensions the "volume" is really area under the curve. This phenomenon is unique to the two-class case. Extending to a three-class case, the ROC space is a hypercube subset of $\mathbb{R}^{3^2-3}=\mathbb{R}^6$, while the CC space is a hypercube inside $\mathbb{R}^3$. The "guessing" observer is a classifier which is a subset of $\mathbb{R}^{3-1}=\mathbb{R}^2$ in ROC space. This clearly has no volume since the linear variety has codimension 4 to the ROC space; however, the guessing observer yields a 3-simplex in CC space, which has a volume of $\frac{1}{3!}=\frac{1}{6}$. These examples can be extend to any n-class system to demonstrate the existence of codimensions >1 which will suffer with similar problems. Further, these examples assume much in the dimensionality and independence of the underlying parameter spaces. Under ideal circumstances where there exist five independent parameters of the classification system, which vary as five of the six conditional probabilities of misclassification, the ROC manifold will be isomorphic to a linear variety in $\mathbb{R}^{3^2-3-1}=\mathbb{R}^5$, which is codimension 1 to ROC space. The second issue to address involves the importance of the parameters a classification system uses. In a three-class example, suppose we have less than five parameters (an occurrence that is acknowledged in[8,9]). Then the codimensionality of the space associated with the ROC manifold will be higher than 1, and no surface can exist. This is a very real possibility! In fact, the dimensionality of the problem has more to do with the underlying parameters of the classification system than with the number of classes, or independent misclassifications.

In this paper, we define the ROC manifold and CC manifold as duals in a given sense. Their analysis is required to describe the classification system. We propose a mathematical definition based on vector space methods to describe both. Unlike previous works, this definition makes no assumption that underlying distributions are known and thus can be utilized when likelihood decision criterion is unavailable. The ROC manifold for n-class classification systems fully describes the system in terms of its misclassifications and, by conjunction, its correct classifications. These manifolds can be used to determine the usefulness of a classification system based on a given performance criterion. We offer the ROC manifold not as a means for finding the optimal classifier through the use of utility or other criteria, but as a means to describe the performance of specific classification systems and to eventually compare performance between systems. Some performance functionals (such as summary statistics) useful for CC manifolds can also be evaluated using the ROC manifold (under certain constraints). Further, the ROC manifold may be computed regardless of the codimension that results from the possible classification systems, that is, directly, without the need to reduce parameters or dimensionality to create a manifold that is codimension 1 to the ROC space. Therefore, the definition of the ROC manifold may subsume previous ROC surface definitions in many cases. Another key difference of the ROC manifold with respect to CC surfaces is that optimal operating parameters may be identified when prior probabilities or costs differ among the various classes. In this paper, we will use the term, *parameter*, to refer to those continuous deterministic quantities that represent different settings for the classification system. These parameters are varied to compare system performance constituting the various points of the ROC manifold. The ROC manifold and CC manifold are paramount to fully evaluating the performances of the classification systems, and herein we endeavor to define them mathematically and describe them in detail.

This paper is constructed as follows. Section 2 outlines the necessary classification system theory. Section 3 defines the ROC manifold and the CC manifold. In this section we observe the relationship between the ROC manifold and the typical ROC curve when only two classes are of interest and between previous surfaces focusing only on correct classifications. We assume underlying distributions are not known and, therefore, likelihood decision criterion is unavailable. We also assume ROCs are invariant with respect to the prevalences of the various classes to be distinguished among, so that the class-conditional probabilities do not change if, or when, prior probabilities change. Section 4 details performance functionals useful for competing two or more classification systems and, specifically, focuses on Bayes cost as a decision criterion. In Section 5, we demonstrate the ROC manifold as useful in finding points of optimal performance defined in terms of the associated misclassification costs and prior probabilities. Using a simple classification system, Section 5 also gives examples that demonstrate the calculation of the ROC manifold and associated optimal points for codimension 1 and higher systems as well as illustrate some properties of this manifold. Further, we demonstrate how the CC manifold fails to identify optimal points when prior probabilities and costs differ, yet the ROC manifold can identify such points. In Section 6, we apply the ROC manifold to real data trying to distinguish between levels of disease progression for chronic allograft nephropathy. Finally, we close with conclusions in Section 6.

2. Classification system theory

The ROC manifold is a mathematical construction comprised a collection of characteristics of a classification system. In this section, we describe the general theory of a classification system and its corresponding collection of characteristics. These characteristics will be used to formally define the ROC manifold in Section 3.

We assume an n-label set, $\mathcal{L}=\{\ell_1,\ell_2,\ldots,\ell_n\}$ where ℓ_k is a generic label and $n \geq 3$. For a 3-label set $\mathcal{L}$, examples include numerical labels such as $\{0,1,-1\}$, nominal (without ordering) categorical labels such as {trains, planes, automobiles} or ordinal (with ordering) categorical labels {normal weight, overweight, obese}. In this paper, we do not assume any ordering on these labels, though such an ordering may exist.

Let $\mathcal{E}$ be a population set of outcomes. We assume there is a truth mapping $\mathbf{T}:\mathcal{E}\to\mathcal{L}$ such that $\mathbf{T}$ partitions the population set with $\{\mathcal{E}_1,\mathcal{E}_2,\ldots,\mathcal{E}_n\}$ where

$$\mathcal{E}_k=\{e\in\mathcal{E}:\mathbf{T}(e)=\ell_k\}$$

for each $k=1,2,\ldots,n$ such that

$$\mathcal{E}_1\cup\mathcal{E}_2\cup\cdots\cup\mathcal{E}_n=\mathcal{E} \quad \text{and} \quad \mathcal{E}_i\cap\mathcal{E}_j=\emptyset \quad \text{for every } i\neq j.$$

Thus,

$$\mathbf{T}(e)=\ell_k \text{ if and only if } e\in\mathcal{E}_k.$$

Let $\mathfrak{E}$ be the σ-algebra of subsets of $\mathcal{E}$ generated by $\mathbf{T}$, then $(\mathcal{E},\mathfrak{E})$ is a measurable space. Let Pr be a probability measure defined on $\mathfrak{E}$, then $(\mathcal{E},\mathfrak{E},\Pr)$ is a probability space.

The truth mapping is the gold standard. In practice, we create a classification system to approximate $\mathbf{T}$. Specifically, we might create a sensor $\mathbf{s}$ to observe the outcome $e\in\mathcal{E}$, that produces a raw datum $d\in\mathcal{D}$, where $\mathcal{D}$ is the (sensor) data set, such that $\mathbf{s}:\mathcal{E}\to\mathcal{D}$. Examples of sensors could be an X-ray device, MRI device, infrared camera, optical camera, camcorder, thermometer, scales, or pressure transducer. Examples of a datum include a chest X-ray, a case of mammogram images, satellite photos, and video. The data set may be too difficult or high dimensional to quantify the attributes for classification, so a feature map $\mathbf{p}$ represents a processor (or feature extractor) that takes a raw datum from $\mathcal{D}$ and produces a refined datum called a feature such that $\mathbf{p}:\mathcal{D}\to\mathcal{F}$, where $\mathcal{F}$ is a feature set. These features are what one (a statistician) commonly refers to as variables, or, as in the chronic allograft nephropathy data presented later, diagnostic markers. Examples of processors that comprise the feature mapping include signal/image processors, signal/image transforms, filters, numerical algorithms, assay kits, and time–frequency transforms. When a nurse computes the body-mass index (BMI) of a patient from his height and weight (the raw data), then the nurse *is* the processor producing the feature (the BMI value). Next we create a classifier $\mathbf{c}$ to assign a label $\ell\in\mathcal{L}$ to each feature $x\in\mathcal{F}$ such that $\mathbf{c}:\mathcal{F}\to\mathcal{L}$. We assume the composition of these mappings, $\mathbf{c}\circ\mathbf{p}\circ\mathbf{s}$, is defined and, hence, yields a new mapping $\mathbf{A}=\mathbf{c}\circ\mathbf{p}\circ\mathbf{s}$. Thus, $\mathbf{A}:\mathcal{E}\to\mathcal{L}$ is defined on some subset of $\mathcal{E}$ (possibly on all of $\mathcal{E}$) and we call the mapping $\mathbf{A}$ a *classification system*. The graphical representation of these mappings is given in the following *diagram*:

$$\mathcal{E}\xrightarrow{\mathbf{s}}\mathcal{D}\xrightarrow{\mathbf{p}}\mathcal{F}\xrightarrow{\mathbf{c}}\mathcal{L}.$$

Suppose there is a parameter $\theta\in\Theta$, a parameter set, that can be varied so that for each $\theta\in\Theta$ the mapping $\mathbf{c}_\theta:\mathcal{F}\to\mathcal{L}$ is another classifier. Note that each parameter refers to a deterministic quantity representing different settings for the classification system. Then the compositions $\mathbf{A}_\theta=\mathbf{c}_\theta\circ\mathbf{p}\circ\mathbf{s}$ yields a collection of systems we denote by $\mathbb{A}=\{\mathbf{A}_\theta:\theta\in\Theta\}$ and call a *family of classification systems* or simply a *classification system family* (CSF). There are other possible ways to construct a CSF. For example, suppose the processor has multiple parameters that can be varied, say $\mathbf{p}_\xi$ for $\xi\in\Xi$, then the CSF is $\mathbb{A}=\{\mathbf{A}_\xi:\xi\in\Xi\}=\{\mathbf{c}\circ\mathbf{p}_\xi\circ\mathbf{s}:\xi\in\Xi\}$. If one can vary the classifier parameters and the processor parameters then another example of a CSF is

$$\mathbb{A}=\{\mathbf{A}_{\theta,\xi}:\theta\in\Theta,\xi\in\Xi\}=\{\mathbf{c}_\theta\circ\mathbf{p}_\xi\circ\mathbf{s}:\theta\in\Theta,\xi\in\Xi\}.$$

In this example, the processor parameter, ξ, may represent different settings such as different feature extractors or neural network weights and the classifier parameter, θ, may represent different model settings or threshold values of interest associated with a statistical model.

Since $\mathcal{L}$ is a finite set and letting $\mathfrak{L}$ denote the power set of $\mathcal{L}$, then $(\mathcal{L},\mathfrak{L})$ is clearly a measurable space. We assume that $\mathbf{A}_\theta$ is a $(\mathfrak{E}-\mathfrak{L})$ measurable mapping for each $\theta\in\Theta$ [16], and, thus, a random mapping (in fact, a stochastic process or random field). Of course, this will be true based upon properties of the individual mappings $\mathbf{c}_\theta$, $\mathbf{p}$, and $\mathbf{s}$. Therefore, appealing to the probability measure space $(\mathcal{E},\mathfrak{E},\Pr)$, we can quantify how well the classification system $\mathbf{A}_\theta$ approximates $\mathbf{T}$ by constructing the ROC manifold. In order to discuss the resulting probability estimates, we recall the definition of a pre-image [16].

Definition 1 (*Pre-image*). Let $\mathcal{X}$ and $\mathcal{Y}$ be nonempty sets. Let the mapping $\mathbf{f}$ take an element from $\mathcal{X}$ and map it into $\mathcal{Y}$, that is, $\mathbf{f}:\mathcal{X}\to\mathcal{Y}$. Given a subset $Y\subset\mathcal{Y}$ we define its *pre-image* with respect to $\mathbf{f}$ to be a subset in $\mathcal{X}$ by

$$\mathbf{f}^{\natural}(Y)=\{x\in\mathcal{X}:\mathbf{f}(x)\in Y\}.$$

Thus, the pre-image of a subset Y in $\mathcal{Y}$ is the collection of all the elements in $\mathcal{X}$ that are mapped by $\mathbf{f}$ into Y.

The pre-image is also called the *inverse image*, although the mapping $\mathbf{f}$ need not be invertible. Because this construction creates a *natural* mapping from subsets of $\mathcal{Y}$ into subsets of $\mathcal{X}$, the natural symbol, $\natural$ (the becuadro), will be used to avoid confusion with the inverse function. Therefore, we write $\mathbf{f}^{\natural}(Y)=X$ and observe that $\mathbf{f}^{\natural}$ maps a set to a set whereas $\mathbf{f}$ maps a point to a point.

Given a classification system $\mathbf{A}$ we write the pre-image of a specific label, $\ell\in\mathcal{L}=\{\ell_1,\ldots,\ell_n\}$ by defining the singleton set $L_\ell=\{\ell\}$, then

$$\mathbf{A}^{\natural}(L_\ell)=\{e\in\mathcal{E}:\mathbf{A}(e)\in L_\ell\}=\{e\in\mathcal{E}:\mathbf{A}(e)=\ell\}.$$

The use of pre-images allows us to take a label and express it in terms of the underlying events.

3. The ROC manifold

3.1. The receiver operating characteristic of a classification system

Assume the classification system $\mathbf{A}:\mathcal{E}\to\mathcal{L}$ is designed to map the outcomes in the event set $\mathcal{E}_i$ to ℓ_i for each $i=1,\ldots,n$. Define the probability of correct classification for a given label ℓ_i of the classification system **A** by the conditional probability

$$P_{i|i}(\mathbf{A}) \equiv \Pr\{\mathbf{A}(e)=\ell_i | e \in \mathcal{E}_i\} = \frac{\Pr(\mathbf{A}^{\natural}(\{\ell_i\}) \cap \mathcal{E}_i)}{\Pr(\mathcal{E}_i)},$$

when $\Pr(\mathcal{E}_i) \neq 0$, and 0 otherwise. We use the notation $P_{i|i}(\mathbf{A})$ to convene the fact that $P_{i|i}$ is a function with the system **A** as its input. The probability that system **A** misclassifies a set of outcomes as label ℓ_i when the outcomes are truly classified as label ℓ_j is

$$P_{i|j}(\mathbf{A}) = \Pr\{\mathbf{A}(e)=\ell_i | e \in \mathcal{E}_j\} = \frac{\Pr(\mathbf{A}^{\natural}(\{\ell_i\}) \cap \mathcal{E}_j)}{\Pr(\mathcal{E}_j)}, \tag{1}$$

when $\Pr(\mathcal{E}j) \neq 0$, and 0 otherwise. There are exactly n conjunctive equations of the system,

$$\sum_{i=1}^{n} P_{i|j}(\mathbf{A}) = 1 \quad \text{for each } j=1,2,\ldots,n \tag{2}$$

[17,18] . The $i|i$ terms are correct classifications, the other $n-1$ terms are the misclassifications of system **A** and, consequently, from Eqs. (2) we have

$$\sum_{i=1, i\neq j}^{n} P_{i|j}(\mathbf{A}) = 1 - P_{j|j}(\mathbf{A}) \quad \text{for each } j=1,2,\ldots,n. \tag{3}$$

For system **A** define the $n \times n$ matrix $\mathtt{P}(\mathbf{A})$ to be the matrix whose i,j entry is the value $P_{i|j}(\mathbf{A})$ for every $i,j \in \{1,\ldots,n\}$, that is,

$$\mathtt{P}(\mathbf{A}) = \begin{bmatrix} P_{1|1}(\mathbf{A}) & P_{1|2}(\mathbf{A}) & P_{1|3}(\mathbf{A}) & \cdots & P_{1|n}(\mathbf{A}) \\ P_{2|1}(\mathbf{A}) & P_{2|2}(\mathbf{A}) & P_{2|3}(\mathbf{A}) & \cdots & P_{2|n}(\mathbf{A}) \\ P_{3|1}(\mathbf{A}) & P_{3|2}(\mathbf{A}) & P_{3|3}(\mathbf{A}) & & P_{3|n}(\mathbf{A}) \\ \vdots & \vdots & & \ddots & \vdots \\ P_{n|1}(\mathbf{A}) & P_{n|2}(\mathbf{A}) & P_{n|3}(\mathbf{A}) & \cdots & P_{n|n}(\mathbf{A}) \end{bmatrix}. \tag{4}$$

This matrix describes the classification information of system **A**. Notice that the diagonal entries of matrix $\mathtt{P}(\mathbf{A})$ are probabilities of correct classification and the off-diagonal entries are probabilities of misclassification. Given the off-diagonal entries, we can compute the diagonal entries by Eq. (3). Therefore, we can represent the probabilities associated with classification by using only the misclassifications of the system, thus losing no classification information. Define $\mathtt{R}(\mathbf{A})$ as the $n \times n$ matrix in which the probabilities of correct classifications are removed,

$$\mathtt{R}(\mathbf{A}) = \begin{bmatrix} 0 & P_{1|2}(\mathbf{A}) & P_{1|3}(\mathbf{A}) & \cdots & P_{1|n}(\mathbf{A}) \\ P_{2|1}(\mathbf{A}) & 0 & P_{2|3}(\mathbf{A}) & \cdots & P_{2|n}(\mathbf{A}) \\ P_{3|1}(\mathbf{A}) & P_{3|2}(\mathbf{A}) & 0 & & P_{3|n}(\mathbf{A}) \\ \vdots & \vdots & & \ddots & \vdots \\ P_{n|1}(\mathbf{A}) & P_{n|2}(\mathbf{A}) & P_{n|3}(\mathbf{A}) & \cdots & 0 \end{bmatrix}. \tag{5}$$

The matrix $\mathtt{R}(\mathbf{A})$ can be derived from $\mathtt{P}(\mathbf{A})$ by

$$\mathtt{R}(\mathbf{A}) = \mathtt{J} \odot \mathtt{P}(\mathbf{A}),$$

where $\odot$ denotes the Hadamard product and matrix $\mathtt{J}$ is defined as

$$\mathtt{J} = \begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & 0 & 1 & \cdots & 1 \\ 1 & 1 & 0 & & 1 \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 0 \end{bmatrix}.$$

The matrix $\mathtt{R}(\mathbf{A})$ is the ROC of the system **A** and is defined as follows.

Definition 2 (*Receiver operating characteristic, ROC*). Let **A** be a classification system. Let $\mathtt{R}(\mathbf{A})$ be a matrix of associated conditional probabilities whose diagonal entries are zero. The matrix $\mathtt{R}(\mathbf{A})$ is the receiver operating characteristic (ROC) of classification system **A**.

Some comments on this definition are given. From Eq. (2) we observe

$$\sum_{i=1, i\neq k}^{n} P_{i|j}(\mathbf{A}) = 1 - P_{k|j}(\mathbf{A}) \quad \text{for each } j=1,2,\ldots,n,$$

where k may or may not equal j and can be different for each j. When $k=j$, a probability of correct classification is removed from the left side. When $k \neq j$ a probability of misclassification is removed. Regardless of the value of k, no classification information of system **A** is lost. The most common application for $k \neq j$ is the two-class case in which the ROC is

$$\mathtt{R}(\mathbf{A}) = \begin{bmatrix} P_{1|1}(\mathbf{A}) & P_{1|2}(\mathbf{A}) \\ 0 & 0 \end{bmatrix}.$$

This ROC is equivalent to the standard false positive-true positive pair of probabilities, i.e., $(P_{1|2}(\mathbf{A}), P_{1|1}(\mathbf{A}))$. The ROC associated with $k=j$ yields

$$\mathtt{R}(\mathbf{A}) = \begin{bmatrix} 0 & P_{1|2}(\mathbf{A}) \\ P_{2|1}(\mathbf{A}) & 0 \end{bmatrix}$$

which produces the probabilities for the two types of misclassifications for the system, i.e., the (α,β) errors. Similarly, the CC manifold [1] is given by

$$\mathtt{R}(\mathbf{A}) = \begin{bmatrix} P_{1|1}(\mathbf{A}) & 0 & \cdots & 0 \\ 0 & P_{2|2}(\mathbf{A}) & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & 0 & 0 & P_{n|n}(\mathbf{A}) \end{bmatrix}. \tag{6}$$

These examples demonstrate that there are other representations of the ROC, however, we will choose to work with the ROC, $\mathtt{R}(\mathbf{A})$ given in Eq. (5). The use of this standard $\mathtt{R}(\mathbf{A})$ will produce what we will call the ROC manifold and yields an intuitive geometric approach when introducing functionals on this ROC manifold.

We note that $\mathtt{Pc}(\mathbf{A}) \equiv \mathtt{I} \odot \mathtt{P}(\mathbf{A})$ yields the matrix of probabilities of correct classifications. While both matrices $\mathtt{Pc}(\mathbf{A})$ and $\mathtt{R}(\mathbf{A})$ have full rank, matrix $\mathtt{R}(\mathbf{A})$ contains n^2-2n more entries (and therefore more information) than matrix $\mathtt{Pc}(\mathbf{A})$. The matrix $\mathtt{Pc}(\mathbf{A})$ is the foundation for the correct classification surfaces described previously in [1–7,10].

3.2. *Construction of the ROC manifold of a classification system family*

Let $\mathcal{R}_n$ denote the set of $n \times n$ matrices $\mathtt{X}$ whose entries lie in [0,1] with zero diagonal entries, that is,

$$\mathcal{R}_n = \{\mathtt{X} = (\mathtt{X}_{i,j}) : \mathtt{X}_{i,j} \in [0,1] \text{ for every } i,j \in \{1,2,\ldots,n\} \quad \text{and} \\ \mathtt{X}_{i,i} = 0 \text{ for all } i=1,2,\ldots,n\}$$

then $\mathtt{R}(\mathbf{A}) \in \mathcal{R}_n$ for every classification system **A**. The set $\mathcal{R}_n$ is the set of ROCs.

Definition 3 (*Error set*). Given a classification system family $\mathbb{A}$, define its error set $\mathcal{E}_{\mathbb{A}}$ to be the set of ROCs

$$\mathcal{E}_{\mathbb{A}} \equiv \{\mathtt{R}(\mathbf{A}) \in \mathcal{R}_n : \mathbf{A} \in \mathbb{A}\}.$$

Clearly $\mathcal{E}_{\mathbb{A}} \subset \mathcal{R}_n$ since the diagonal entries are always zero. This error set could be as large as the family $\mathbb{A}$, however, we seek a subset of $\mathcal{E}_{\mathbb{A}}$ such that its members lie "closest" to the origin as we define next.

Definition 4 (*ROC function*). Given a classification system family $\mathbb{A}$, define its ROC function $\Upsilon_{\mathbb{A}}$ to be, for every $\mathtt{X} \in \mathcal{R}_n$

$$\Upsilon_{\mathbb{A}}(\mathtt{X}) = \begin{cases} \text{smallest } \alpha \geq 0 & \text{such that } \alpha\mathtt{X} \in \mathcal{E}_{\mathbb{A}} \\ \infty & \text{if } \alpha\mathtt{X} \notin \mathcal{E}_{\mathbb{A}} \text{ for all } \alpha \geq 0 \end{cases}$$
$$= \min\{\alpha \in [0,\infty) : \alpha\mathtt{X} \in \mathcal{E}_{\mathbb{A}}\}.$$

Therefore, $\Upsilon_{\mathbb{A}} : \mathcal{R}_n \to [0,\infty]$.

This definition is motivated by the Minkowski functional [19,20]. An equivalent definition of $\Upsilon_{\mathbb{A}}$, useful for computations, is given in the following theorem using the alignment concept [19]. We say two matrices $\mathtt{X}$ and $\mathtt{Y}$ are aligned if

$$\langle \mathtt{X},\mathtt{Y} \rangle = \|\mathtt{X}\|\|\mathtt{Y}\|,$$

where $\langle \cdot , \cdot \rangle$ is the inner product that generates the Frobenius norm $\| \cdot \|$. Basically, $\mathtt{X}$ and $\mathtt{Y}$ "point" in the same direction, but may not have the same "length".

Theorem 1. *Given a classification system family* $\mathbb{A}$, *for every* $0 \neq \mathtt{X} \in \mathcal{R}_n$ *define the set of ROCs aligned with* $\mathtt{X}$ *to be*

$$\mathcal{A}_{\mathbb{A}}(\mathtt{X}) \equiv \{\mathtt{Y} \in \mathcal{E}_{\mathbb{A}} : \langle \mathtt{Y},\mathtt{X} \rangle = \|\mathtt{Y}\|\|\mathtt{X}\|\}.$$

Then, the ROC function $\Upsilon_{\mathbb{A}}$ *is*

$$\Upsilon_{\mathbb{A}}(\mathtt{X}) = \begin{cases} \dfrac{1}{\|\mathtt{X}\|} \min\limits_{\mathtt{Y} \in \mathcal{A}_{\mathbb{A}}(\mathtt{X})} \|\mathtt{Y}\| & \text{if } \mathcal{A}_{\mathbb{A}}(X) \neq \emptyset, \\ \infty & \text{if } \mathcal{A}_{\mathbb{A}}(X) = \emptyset. \end{cases}$$

Proof. Let $0 \neq \mathtt{X} \in \mathcal{R}_n$. For every $\alpha > 0$ the matrix $\mathtt{Y} = \alpha\mathtt{X}$ is aligned with $\mathtt{X}$, that is,

$$\langle \mathtt{Y},\mathtt{X} \rangle = \langle \alpha\mathtt{X},\mathtt{X} \rangle = \alpha\|\mathtt{X}\|^2 = \alpha\|\mathtt{X}\|\|\mathtt{X}\| = \|\alpha\mathtt{X}\|\|\mathtt{X}\| = \|\mathtt{Y}\|\|\mathtt{X}\|.$$

Since,

$$\|\mathtt{Y}\| = \|\alpha\mathtt{X}\| = \alpha\|\mathtt{X}\|$$

then

$$\alpha = \frac{\|\mathtt{Y}\|}{\|\mathtt{X}\|}.$$

Therefore,

$$\begin{aligned} \{\alpha > 0 : \alpha\mathtt{X} \in \mathcal{E}_{\mathbb{A}}\} &= \left\{\frac{\|\mathtt{Y}\|}{\|\mathtt{X}\|} : \mathtt{Y} \in \mathcal{E}_{\mathbb{A}} \text{ and } \mathtt{Y} \text{ is aligned with } \mathtt{X}\right\} \\ &= \frac{1}{\|\mathtt{X}\|}\{\|\mathtt{Y}\| : \mathtt{Y} \in \mathcal{E}_{\mathbb{A}} \text{ and } \mathtt{Y} \text{ is aligned with } \mathtt{X}\} \\ &= \frac{1}{\|\mathtt{X}\|}\{\|\mathtt{Y}\| : \mathtt{Y} \in \mathcal{A}_{\mathbb{A}}(\mathtt{X})\}. \end{aligned}$$

Hence,

$$\begin{aligned} \Upsilon_{\mathbb{A}}(\mathtt{X}) &= \min\{\alpha > 0 : \alpha\mathtt{X} \in \mathcal{E}_{\mathbb{A}}\} \\ &= \frac{1}{\|\mathtt{X}\|} \min_{\mathtt{Y} \in \mathcal{A}_{\mathbb{A}}(\mathtt{X})} \|\mathtt{Y}\|. \end{aligned}$$

If the set $\mathcal{A}_{\mathbb{A}}(\mathtt{X})$ is empty then the minimum of $\mathcal{A}_{\mathbb{A}}(\mathtt{X})$ is defined to be $+\infty$. □

Here we have a calculation which takes the minimum "Euclidean" distance (i.e., the Frobenius matrix norm) of all matrices aligned in the same direction as $\mathtt{X}$. The inner product $\langle \mathtt{X},\mathtt{Y} \rangle$ is computed by trace($\mathtt{X}^{\mathrm{T}}\mathtt{Y}$).

Definition 5 (*ROC manifold*). Given a classification system family $\mathbb{A}$ define its ROC manifold $\mathcal{M}_{\mathbb{A}}$ to be

$$\mathcal{M}_{\mathbb{A}} = \{\mathtt{X} \in \mathcal{R}_n : \Upsilon_{\mathbb{A}}(\mathtt{X}) = 1\}.$$

The intent is to express the ROC manifold in terms of those minimal errors so that, by property (3), the correct classifications are maximized. Therefore, observe that the systems $\mathbf{A} \in \mathbb{A}$ such that $\Upsilon_{\mathbb{A}}(\mathtt{R}(\mathbf{A})) = 1$ are the systems of interest and those corresponding matrices $\mathtt{R}(\mathbf{A})$ form the ROC manifold. The definition of a manifold is the following. "A topological m-manifold is a Hausdorff topological space for which each point has a neighborhood homeomorphic to an open subset of $\mathbb{R}^m$ (typically the unit ball)" [21].

4. Performance of a classification system

How does one quantify the approximation of a classification system $\mathbf{A}$ to the truth system $\mathbf{T}$, especially when we do not know $\mathbf{T}$? We construct a summary functional ρ to quantify the approximation to $\mathbf{T}$. Thus, we require $\rho(\mathbf{A})$ to be a non-negative real number, and for definiteness, assume a smaller value of $\rho(\mathbf{A})$ implies a better approximation. So, if $\rho(\mathbf{A}) < \rho(\mathbf{B})$ then system $\mathbf{A}$ approximates $\mathbf{T}$ better than $\mathbf{B}$. The value $\rho(\mathbf{A})$ is called the *performance* of $\mathbf{A}$, and the functional ρ is called a *system performance functional*. A system performance functional ρ induces a *family performance functional* ϱ on a classification system family $\mathbb{A}$ by defining

$$\varrho(\mathbb{A}) = \min_{\mathbf{A} \in \mathbb{A}} \rho(\mathbf{A})$$

and, in the case $\mathbb{A} = \{\mathbf{A}_\theta : \theta \in \Theta\}$, then

$$\varrho(\mathbb{A}) = \min_{\mathbf{A} \in \mathbb{A}} \rho(\mathbf{A}) = \min_{\theta \in \Theta} \rho(\mathbf{A}_\theta).$$

A performance functional will necessarily involve the misclassifications of the system. Given ROC $\mathtt{R}(\mathbf{A})$, then for any function $\varphi : \mathcal{R}_n \to \mathbb{R}^+ = [0,\infty)$, a system performance functional ρ is created by the composition

$$\rho(\mathbf{A}) = \varphi(\mathtt{R}(\mathbf{A}))$$

and, consequently,

$$\varrho(\mathbb{A}) = \min_{\mathbf{A} \in \mathbb{A}} \rho(\mathbf{A}) = \min_{\mathbf{A} \in \mathbb{A}} \varphi(\mathtt{R}(\mathbf{A})) = \min_{\mathtt{X} \in \mathcal{M}_{\mathbb{A}}} \varphi(\mathtt{X}).$$

4.1. Comparing classification system families

The family performance functional can be used to compare different CSFs to determine which classifies better (approximates $\mathbf{T}$ better). Thus, given two CSFs $\mathbb{A}$ and $\mathbb{B}$, we wish to determine which CSF is "better" with respect to the family performance functional ϱ. If

$$\varrho(\mathbb{A}) \leq \varrho(\mathbb{B})$$

then we say $\mathbb{A}$ is better than or equal to $\mathbb{B}$ with respect to ϱ. (Recall a smaller value of $\varrho(\mathbb{A})$ means better performance.) This induces a partial ordering, $\overset{\rho}{\preceq}$, on the collection of CSFs, and hence, we write

$$\mathbb{A} \overset{\varrho}{\succcurlyeq} \mathbb{B}.$$

Now we develop a specific functional, Bayes cost. We choose this functional because it will allow us to compare CSFs regardless of our prior probabilities or costs of misclassifications; that is, we can determine the CSF that minimizes errors subject to the underlying prior probabilities and misclassification costs.

4.2. Expected cost

Let $p_j \equiv \Pr(\mathcal{E}_j)$, the prior probability of class j, and $c_{i|j} \geq 0$, the cost of classifying outcome $e \in \mathcal{E}_j$ as label ℓ_i. This is a correct classification cost when i=j and a misclassification cost when $i \neq j$. Now we can define the system performance functional (expected cost) by

$$\rho(\mathbf{A}_\theta)=\varphi(\mathrm{R}(\mathbf{A}_\theta))=\sum_{j=1}^{n}\sum_{i=1,i\neq j}^{n} c_{i|j}p_jP_{i|j}(\mathbf{A}_\theta). \tag{7}$$

Now consider a couple of options with this equation. First option could be allowing $c_{i|j}$ to equal 0 whenever $i\neq j$. This would allow a functional over the CC manifold. Indeed, this expected cost performance functional is what we would like to be maximized. On the other hand, we could choose instead for $c_{i|j}$ to equal 0 whenever $i=j$, giving risk, an expected cost performance functional we would like to minimize. In this second case

$$\rho(\mathbf{A}_\theta)=\varphi(\mathrm{R}(\mathbf{A}_\theta))=\sum_{j=1}^{n}\sum_{i=j}^{n} c_{i|j}p_jP_{i|j}(\mathbf{A}_\theta)=\langle \Gamma,\mathrm{R}(\mathbf{A}_\theta)\rangle, \tag{8}$$

where the matrix $\Gamma=\mathrm{C}\cdot\mathrm{diag}(\mathrm{p})$ for the cost matrix C and the diagonal matrix, $\mathrm{diag}(\mathrm{p})$ of prior probabilities. Thus, the family performance functional

$$\varrho(\mathbb{A})=\min_{\theta\in\Theta}\langle \Gamma,\mathrm{R}(\mathbf{A}_\theta)\rangle$$

is the Bayes cost of the CSF $\mathbb{A}$. It is reasonable to use this functional when good estimates of the priors and costs are available. Note that when each $c_{i|j}$ is equal to 0 for $i=j$, the inner product in Eq. (8) is equivalent to $\langle \Gamma,\mathrm{P}(\mathbf{A}_\theta)\rangle$ and thus can be minimized using this product as well.

5. Examples

This section is divided into three parts with which to illustrate properties of the ROC manifold. First, in Section 5.1, we show generally how to compute the ROC manifold for a specific CSF and, using Bayes cost, how to compare between candidate classification systems. To aid computations, we assume distributional information is known, though it need not be. Within this part, we offer example 1 and example 2 which compete systems based on their optimal parameter settings using equal (example 1) and unequal (example 2) costs and prior probabilities. Section 5.2 demonstrates the dual problem through the computation of the CC manifold from the distributions and classifiers suggested in Section 5.1. The CC manifold is computed and examples 1 (equal costs and priors) and 2 (unequal costs and priors) are again given. Now, identical solutions are shown between the CC manifold in Example 1 of Section 5.2 and the ROC manifold in Example 1 of Section 5.1. However, we show mathematically the inability of finding a solution from the CC manifold of Example 2 in this Section 5.2. Finally, in Section 5.3, we demonstrate computation of the ROC manifold when the codimension is greater than 1. To aid computation, we again assume distributional information is known and suggest a classifier different than that presented in previous examples. We conclude this example by computing the optimal point via Bayes cost and suggest the optimal operating parameters for that specific classifier.

5.1. Computation and classification system comparisons using the ROC manifold

To illustrate the ROC manifold of a CSF, different performances, and different comparison of CSFs, we pose an example of a 3-label classification problem. Let the label set $\mathcal{L}=\{\ell_1,\ell_2,\ell_3\}$ have generic labels where the indices 1, 2 and 3 do not necessarily imply any ordering. Assume that the sensor $\mathbf{s}$ and processor $\mathbf{p}$ mapped the disjoint events $\mathcal{E}_1,\mathcal{E}_2,\mathcal{E}_3$ into the feature set $\mathcal{F}=\mathbb{R}$ and produced sets that were not disjoint but normally distributed with density distributions given by

Label	Center	Variance	Density
ℓ_1	-1	0.5	$f_1(x)=\dfrac{1}{0.5\sqrt{\pi}}\exp\left(-\left(\dfrac{x+1}{0.5}\right)^2\right)$
ℓ_2	0	1	$f_2(x)=\dfrac{1}{\sqrt{\pi}}\exp(-x^2)$
ℓ_3	1	2	$f_3(x)=\dfrac{1}{2\sqrt{\pi}}\exp\left(-\left(\dfrac{x-1}{2}\right)^2\right)$

Here, $x\in\mathbb{R}$ is a feature. These density distributions are graphed in Fig. 1.

Define the classifier

$$\mathbf{a}_{\theta_1,\theta_2}(x)=\begin{cases}\ell_1 & \text{for } -\infty<x<\theta_1,\\ \ell_2 & \text{for } \theta_1\le x<\theta_2,\\ \ell_3 & \text{for } \theta_2\le x<\infty\end{cases}$$

for the two parameters $\theta_1,\theta_2\in[-5,5]$ such that $\theta_1\le\theta_2$. These parameters will act as threshold cut points for the classifier $\mathbf{a}_{\theta_1,\theta_2}$. For brevity of notation we write

$$\Theta=[-5,5]_+^2\equiv\{\boldsymbol{\theta}=(\theta_1,\theta_2)\in[-5,5]\times[-5,5]:\theta_1\le\theta_2\}.$$

The composition $\mathbf{a}_{\theta_1,\theta_2}\circ\mathbf{p}\circ\mathbf{s}$ yields a classification system $\mathbf{A}_{\boldsymbol{\theta}}=\mathbf{A}_{\theta_1,\theta_2}\equiv\mathbf{a}_{\theta_1,\theta_2}\circ\mathbf{p}\circ\mathbf{s}$ for each $(\theta_1,\theta_2)\in[-5,5]_+^2$. The resulting CSF is

$$\mathbb{A}=\{\mathbf{A}_{\theta_1,\theta_2}:(\theta_1,\theta_2)\in[-5,5]_+^2\}.$$

For $\theta_1\le\theta_2$ the conditional probabilities are computed by

$$P_{1|j}(\mathbf{A}_{\boldsymbol{\theta}})=\int_{-\infty}^{\theta_1}f_j(x)\,dx,$$

$$P_{2|j}(\mathbf{A}_{\boldsymbol{\theta}})=\int_{\theta_1}^{\theta_2}f_j(x)\,dx,$$

$$P_{3|j}(\mathbf{A}_{\boldsymbol{\theta}})=\int_{\theta_2}^{\infty}f_j(x)\,dx$$

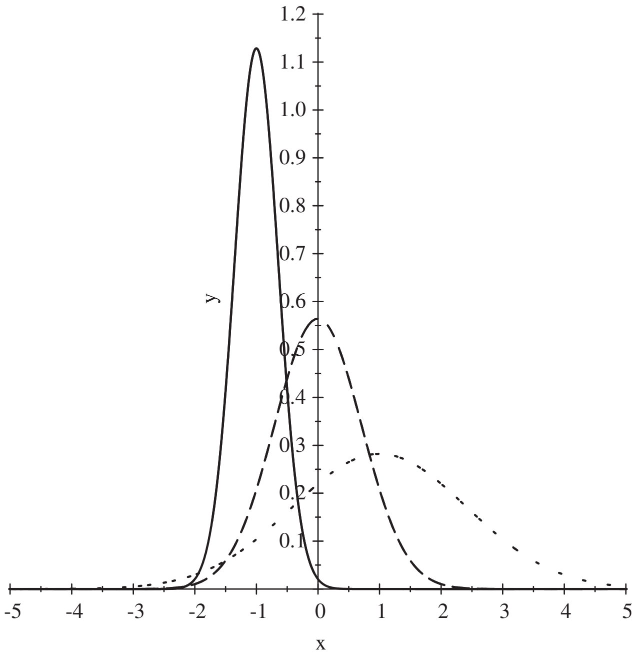

Fig. 1. The plots of the one-dimensional density functions f_1 (solid), f_2 (dashes), and f_3 (dots).

for $j=1,2,3$, and the associated matrix of conditional probabilities from Eq. (4) is

$$\mathrm{P}(\mathbf{A}_\theta)=\frac{1}{2}\begin{bmatrix} 2-\mathrm{erfc}(2\theta_1+2) & \mathrm{erf}(\theta_1)+1 & \mathrm{erf}\left(\frac{\theta_1-1}{2}\right)+1 \\ \mathrm{erfc}(2\theta_1+2)-\mathrm{erfc}(2\theta_2+2) & \mathrm{erf}(\theta_2)-\mathrm{erf}(\theta_1) & \mathrm{erf}\left(\frac{\theta_2-1}{2}\right)-\mathrm{erf}\left(\frac{\theta_1-1}{2}\right) \\ \mathrm{erfc}(2\theta_2+2) & 1-\mathrm{erf}(\theta_2) & 1-\mathrm{erf}\left(\frac{\theta_2-1}{2}\right)\end{bmatrix}.$$

Since there are two free parameters θ_1, θ_2, the ROC manifold is a two-dimensional manifold lying in the six-dimensional set $\mathcal{R}_3$ (observe that $\mathcal{R}_3$ is isomorphic to $[0,1]^6$). Whereas this manifold cannot be visually graphed, the performance functionals can still be applied.

Example 1. Consider the Bayes cost performance functional with equal prior probabilities $p_j=\Pr(\mathcal{E}_j)=\frac{1}{3}$ and equal costs $c_{i|j}=1$ for all $i\neq j$, then the matrix Γ given in Eq. (8) is

$$\Gamma=\begin{bmatrix}0 & c_{1|2} & c_{1|3}\\ c_{2|1} & 0 & c_{2|3}\\ c_{3|1} & c_{3|2} & 0\end{bmatrix}\begin{bmatrix}p_1 & 0 & 0\\ 0 & p_2 & 0\\ 0 & 0 & p_3\end{bmatrix}=\begin{bmatrix}0 & 1 & 1\\ 1 & 0 & 1\\ 1 & 1 & 0\end{bmatrix}\begin{bmatrix}\frac{1}{3} & 0 & 0\\ 0 & \frac{1}{3} & 0\\ 0 & 0 & \frac{1}{3}\end{bmatrix}$$

$$=\begin{bmatrix}0 & \frac{1}{3} & \frac{1}{3}\\ \frac{1}{3} & 0 & \frac{1}{3}\\ \frac{1}{3} & \frac{1}{3} & 0\end{bmatrix} \tag{9}$$

and the inner product yields

$$\langle \Gamma,\mathrm{R}(\mathbf{A}_\theta)\rangle$$
$$=\frac{1}{2}\mathrm{trace}\begin{bmatrix}0 & \frac{1}{3} & \frac{1}{3}\\ \frac{1}{3} & 0 & \frac{1}{3}\\ \frac{1}{3} & \frac{1}{3} & 0\end{bmatrix}^{\mathrm{T}}$$
$$\times\begin{bmatrix} 0 & \mathrm{erf}(\theta_1)+1 & \mathrm{erf}\left(\frac{\theta_1-1}{2}\right)+1 \\ \mathrm{erfc}(2\theta_1+2)-\mathrm{erfc}(2\theta_2+2) & 0 & \mathrm{erf}\left(\frac{\theta_2-1}{2}\right)-\mathrm{erf}\left(\frac{\theta_1-1}{2}\right) \\ \mathrm{erfc}(2\theta_2+2) & 1-\mathrm{erf}(\theta_2) & 0\end{bmatrix}$$
$$=\frac{1}{6}\left[\mathrm{erfc}(2\theta_1+2)+\mathrm{erf}(\theta_1)-\mathrm{erf}(\theta_2)+\mathrm{erf}\left(\frac{\theta_2-1}{2}\right)+3\right].$$

Therefore,

$$\varrho(\mathbb{A})=\min_{\theta\in\Theta}\langle \Gamma,\mathrm{R}(\mathbf{A}_\theta)\rangle$$
$$=\frac{1}{6}\min_{\theta\in\Theta}\left[\mathrm{erfc}(2\theta_1+2)+\mathrm{erf}(\theta_1)-\mathrm{erf}(\theta_2)+\mathrm{erf}\left(\frac{\theta_2-1}{2}\right)+3\right]$$
$$=0.297$$

occurring at $(\theta_1^*,\theta_2^*)=(-0.511,0.837)$. These optimal cut points are plotted as dashed vertical lines in Fig. 2.

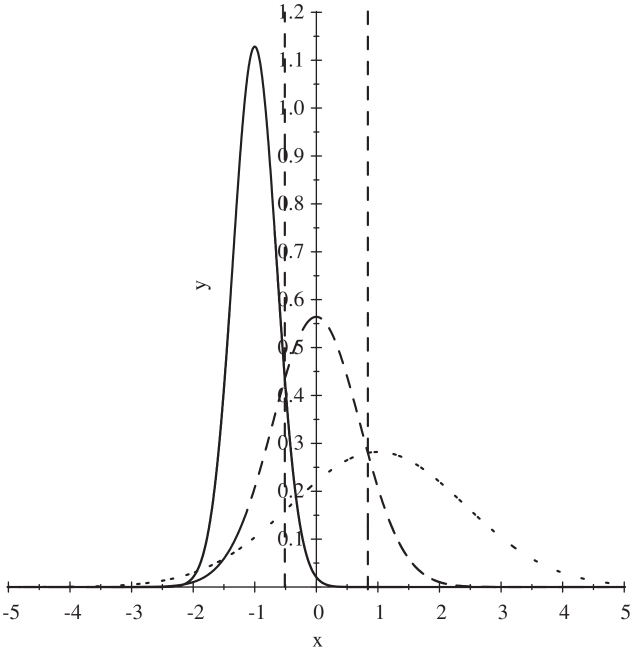

Fig. 2. The plot of the three density functions and the optimal cut points $(\theta_1^*,\theta_2^*)=(-0.511,0.837)$ as dashed vertical lines for Example 1 (equal costs and equal prior probabilities).

Example 2. If we choose a Bayes performance functional with differing prior probabilities and costs to be

$$\Gamma=\begin{bmatrix}0 & c_{1|2} & c_{1|3}\\ c_{2|1} & 0 & c_{2|3}\\ c_{3|1} & c_{3|2} & 0\end{bmatrix}\begin{bmatrix}p_1 & 0 & 0\\ 0 & p_2 & 0\\ 0 & 0 & p_3\end{bmatrix}$$

$$=\begin{bmatrix}0 & 1 & 3\\ 2 & 0 & 2\\ 1 & 3 & 0\end{bmatrix}\begin{bmatrix}\frac{1}{2} & 0 & 0\\ 0 & \frac{1}{3} & 0\\ 0 & 0 & \frac{1}{6}\end{bmatrix}=\begin{bmatrix}0 & \frac{1}{3} & \frac{1}{2}\\ 1 & 0 & \frac{1}{3}\\ \frac{1}{2} & 1 & 0\end{bmatrix} \tag{10}$$

then the inner product is

$$\langle \Gamma,\mathrm{R}(\mathbf{A}_\theta)\rangle$$
$$=\frac{1}{2}\mathrm{trace}\begin{bmatrix}0 & \frac{1}{3} & \frac{1}{2}\\ 1 & 0 & \frac{1}{3}\\ \frac{1}{2} & 1 & 0\end{bmatrix}^{\mathrm{T}}$$
$$\times\begin{bmatrix} 0 & \mathrm{erf}(\theta_1)+1 & \mathrm{erf}\left(\frac{\theta_1-1}{2}\right)+1 \\ \mathrm{erfc}(2\theta_1+2)-\mathrm{erfc}(2\theta_2+2) & 0 & \mathrm{erf}\left(\frac{\theta_2-1}{2}\right)-\mathrm{erf}\left(\frac{\theta_1-1}{2}\right) \\ \mathrm{erfc}(2\theta_2+2) & 1-\mathrm{erf}(\theta_2) & 0\end{bmatrix}$$
$$=\frac{1}{12}\left(\mathrm{erf}\left(\frac{\theta_1-1}{2}\right)+2\mathrm{erf}(\theta_1)+6\mathrm{erfc}(2\theta_1+2)\right.$$
$$\left.+2\mathrm{erf}\left(\frac{\theta_2-1}{2}\right)-6\mathrm{erf}(\theta_2)-3\mathrm{erfc}(2\theta_2+2)+11\right)$$

and the Bayes cost is

$$\varrho(\mathbb{A})=\min_{\theta\in\Theta}\langle \Gamma,\mathrm{R}(\mathbf{A}_\theta)\rangle$$
$$=\frac{1}{12}\min_{\theta\in\Theta}\left(\mathrm{erf}\left(\frac{\theta_1-1}{2}\right)+2\mathrm{erf}(\theta_1)+6\mathrm{erfc}(2\theta_1+2)\right.$$
$$\left.+2\mathrm{erf}\left(\frac{\theta_2-1}{2}\right)-6\mathrm{erf}(\theta_2)-3\mathrm{erfc}(2\theta_2+2)+11\right)=0.576.$$

The optimal parameters are $(\theta_1^*,\theta_2^*)=(-0.746,-0.125)$ with minimum value of 0.576. These cut points are plotted as dashed vertical lines in Fig. 3. With unequal values, the optimal cut points do not lie in a visually intuitive location as they do in Fig. 2. Thus, using visual inspection to deduce optimal points for lower

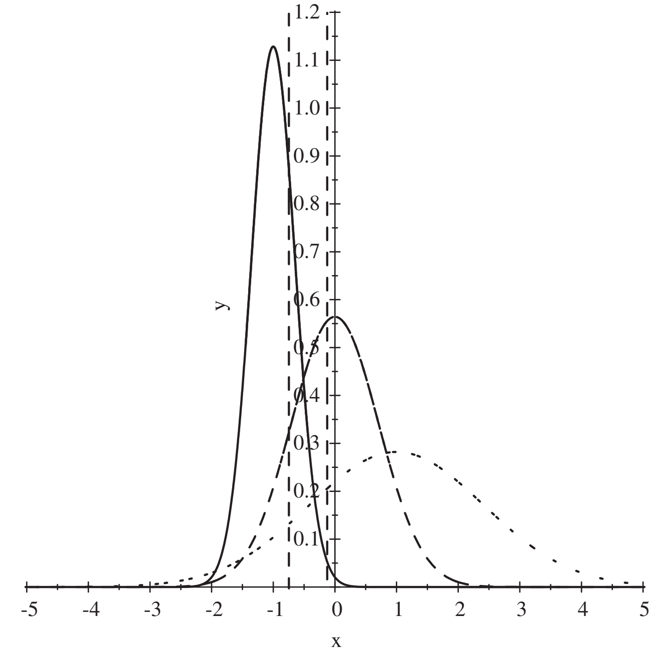

Fig. 3. The plot of the three density functions and the optimal cut points $(\theta_1^*,\theta_2^*)=(-0.746,-0.125)$ as dashed vertical lines for Example 2 (unequal costs and unequal prior probabilities).

dimensional (graphable) systems, without first conducting supporting computations, will not necessarily produce optimal cut points any time that the prior probabilities or the costs are unequal.

5.2. Computation and comparisons with the correct classification (CC) method

Now we compare with the CC method [1] for determining the optimal cut points. This method maximizes the probabilities of correct classifications over the CC manifold. Specifically, the functional for the correct classification is

$$\varrho_{\mathrm{cc}}(\mathbb{A})=\max_{\boldsymbol{\theta}\in\Theta}\sum_{i=1}^{n} d_{i|i}p_iP_{i|i}(\mathbf{A}_{\boldsymbol{\theta}}),$$

where the costs $d_{i|i}\geq 0$ for $i=1,2,\ldots,n$ represent the costs (or weights) for the ith correct classification. This expression is equivalent to that of expected cost in Eq. (7) of Section 4.2 in which $c_{i|j}$ equals 0 whenever $i\neq j$ and produces a functional defined on the CC manifold. Indeed, this expected cost performance functional is what we would like to be maximized. Since

$$\sum_{i=1}^{3}P_{i|j}(\mathbf{A}_\theta)=1$$

then by Eq. (3) we have

$$P_{1|1}(\mathbf{A}_\theta)=1-P_{2|1}(\mathbf{A}_\theta)-P_{3|1}(\mathbf{A}_\theta),$$

$$P_{2|2}(\mathbf{A}_\theta)=1-P_{1|2}(\mathbf{A}_\theta)-P_{3|2}(\mathbf{A}_\theta),$$

$$P_{3|3}(\mathbf{A}_\theta)=1-P_{1|3}(\mathbf{A}_\theta)-P_{2|3}(\mathbf{A}_\theta).$$

Observe that

$$\begin{aligned}
&\sum_{i=1}^{3}d_{i|i}p_iP_{i|i}(\mathbf{A}_\theta)\\
&\quad=d_{1|1}p_1P_{1|1}(\mathbf{A}_\theta)+d_{2|2}p_2P_{2|2}(\mathbf{A}_\theta)+d_{3|3}p_3P_{3|3}(\mathbf{A}_\theta)\\
&\quad=d_{1|1}p_1[1-P_{2|1}(\mathbf{A}_\theta)-P_{3|1}(\mathbf{A}_\theta)]+d_{2|2}p_2[1-P_{1|2}(\mathbf{A}_\theta)-P_{3|2}(\mathbf{A}_\theta)]\\
&\qquad+d_{3|3}p_3[1-P_{1|3}(\mathbf{A}_\theta)-P_{2|3}(\mathbf{A}_\theta)]\\
&\quad=\sum_{i=1}^{3}d_{i|i}p_i-d_{1|1}p_1[P_{2|1}(\mathbf{A}_\theta)+P_{3|1}(\mathbf{A}_\theta)]-d_{2|2}p_2[P_{1|2}(\mathbf{A}_\theta)+P_{3|2}(\mathbf{A}_\theta)]\\
&\qquad-d_{3|3}p_3[P_{1|3}(\mathbf{A}_\theta)+P_{2|3}(\mathbf{A}_\theta)]\\
&\quad=\sum_{i=1}^{3}d_{i|i}p_i-\mathrm{trace}\begin{bmatrix}0 & d_{2|2}p_2 & d_{3|3}p_3\\ d_{1|1}p_1 & 0 & d_{3|3}p_3\\ d_{1|1}p_1 & d_{2|2}p_2 & 0\end{bmatrix}^{\mathsf{T}}\\
&\qquad\times\begin{bmatrix}0 & P_{1|2}(\mathbf{A}_\theta) & P_{1|3}(\mathbf{A}_\theta)\\ P_{2|1}(\mathbf{A}_\theta) & 0 & P_{2|3}(\mathbf{A}_\theta)\\ P_{3|1}(\mathbf{A}_\theta) & P_{3|2}(\mathbf{A}_\theta) & 0\end{bmatrix}\\
&\quad=\sum_{i=1}^{3}d_{i|i}p_i-\mathrm{trace}\varGamma^{\mathsf{T}}\mathrm{R}(\mathbf{A}_\theta).
\end{aligned}$$

Thus, maximizing over Θ yields

$$\begin{aligned}
\varrho_{\mathrm{cc}}(\mathbb{A})&=\max_{\boldsymbol{\theta}\in\Theta}\sum_{i=1}^{n} d_{i|i}p_iP_{i|i}(\mathbf{A}_\theta)\\
&=\max_{\boldsymbol{\theta}\in\Theta}\left(\sum_{i=1}^{3}d_{i|i}p_i-\mathrm{trace}\ \varGamma^{\mathsf{T}}\mathrm{R}(\mathbf{A}_\theta)\right)\\
&=\sum_{i=1}^{3}d_{i|i}p_i-\min_{\boldsymbol{\theta}\in\Theta}\mathrm{trace}\ \varGamma^{\mathsf{T}}\mathrm{R}(\mathbf{A}_\theta).
\end{aligned}$$

The matrix $\varGamma$ is

$$\varGamma=\begin{bmatrix}0 & d_{2|2}p_2 & d_{3|3}p_3\\ d_{1|1}p_1 & 0 & d_{3|3}p_3\\ d_{1|1}p_1 & d_{2|2}p_2 & 0\end{bmatrix}=\begin{bmatrix}0 & d_{2|2} & d_{3|3}\\ d_{1|1} & 0 & d_{3|3}\\ d_{1|1} & d_{2|2} & 0\end{bmatrix}\begin{bmatrix}p_1 & 0 & 0\\ 0 & p_2 & 0\\ 0 & 0 & p_3\end{bmatrix}.$$

For this special form of $\varGamma$ we get

$$\varrho_{\mathrm{cc}}(\mathbb{A})=\sum_{i=1}^{3}d_{i|i}p_i-\varrho(\mathbb{A}). \tag{11}$$

Example 1. Assume equal prior probabilities $p_i=\Pr(\mathcal{E}_i)=\frac{1}{3}$ and equal costs $d_{i|i}=1$ for all $i=1,2,3$. Then

$$\begin{aligned}
\varrho_{\mathrm{cc}}(\mathbb{A})&=\max_{\boldsymbol{\theta}\in\Theta}\sum_{i=1}^{3} d_{i|i}p_iP_{i|i}(\mathbf{A}_\theta)\\
&=\max_{\boldsymbol{\theta}\in\Theta}\frac{1}{3}[P_{1|1}(\mathbf{A}_\theta)+P_{2|2}(\mathbf{A}_\theta)+P_{3|3}(\mathbf{A}_\theta)]\\
&=\max_{\boldsymbol{\theta}\in\Theta}\frac{1}{6}\left[3-\mathrm{erfc}(2\theta_1+2)+\mathrm{erf}(\theta_2)-\mathrm{erf}(\theta_1)-\mathrm{erf}\left(\frac{\theta_2-1}{2}\right)\right]\\
&=0.703
\end{aligned}$$

occurring for $(\theta_1^*,\theta_2^*)=(-0.511,0.837)$ which are the same cut points determined from the ROC manifold using equal prior probabilities and costs. Thus, the Bayes cost functional ϱ and the CC functional ϱ_{cc} identify the same cut points for equal priors and costs. Observe the $\varGamma$ matrix given in Eq. (9) is of the form

$$\varGamma=\begin{bmatrix}0 & b & c\\ a & 0 & c\\ a & b & 0\end{bmatrix}\begin{bmatrix}p_1 & 0 & 0\\ 0 & p_2 & 0\\ 0 & 0 & p_3\end{bmatrix}. \tag{12}$$

Hence,

$$\varrho_{\mathrm{cc}}(\mathbb{A})=1-\varrho(\mathbb{A}). \tag{13}$$

Example 2. Assuming different prior probabilities and different misclassification costs yield the Bayes cost performance functional found in Example 2 of Section 5.1 for the ROC manifold. The optimal parameters are $(\theta_1^*,\theta_2^*)=(-0.746,-0.125)$ with performance value of $\varrho(\mathbb{A})=0.576$. However, since the Γ matrix, given in Eq. (10), is not of the special form as in Eq. (12) then Eq. (11) for the CC method does not hold true. To prove this point, we seek correct classification costs $d_{i|i}\geq 0$ that yields the same optimal parameters, $(\theta_1^*,\theta_2^*)=(-0.746,-0.125)$. Let $d_{1|1},d_{2|2}$ and $d_{3|3}$ be unspecified. Then

$$\begin{aligned}
&\varrho_{cc}(\mathbb{A})\\
&\quad=\max_{\theta\in\Theta}\sum_{i=1}^{3}d_{i|i}p_iP_{i|i}(\mathbf{A}_\theta)\\
&\quad=\max_{\theta\in\Theta}\left(\frac{1}{2}d_{1|1}P_{1|1}(\mathbf{A}_\theta)+\frac{1}{3}d_{2|2}P_{2|2}(\mathbf{A}_\theta)+\frac{1}{6}d_{3|3}P_{3|3}(\mathbf{A}_\theta)\right)\\
&\quad=\max_{\theta\in\Theta}\left(\frac{1}{2}d_{1|1}\left[\frac{2-\operatorname{erfc}(2\theta_1+2)}{2}\right]+\frac{1}{3}d_{2|2}\left[\frac{\operatorname{erf}(\theta_2)-\operatorname{erf}(\theta_1)}{2}\right]\right.\\
&\qquad\left.+\frac{1}{6}d_{3|3}\left[\frac{1-\operatorname{erf}\left(\frac{\theta_2-1}{2}\right)}{2}\right]\right).
\end{aligned}$$

classifications, and ultimately, upon the ROC manifold. In special cases, the CC method and Bayes cost applied to the ROC manifold will yield the same result as illustrated in our first example with equal priors and costs. However, there are other cases for which the CC method cannot always produce an answer, yet the ROC manifold can, i.e., when priors and costs differ.

5.3. The ROC manifold with higher codimension

We illustrate a ROC manifold of a CSF that is not codimension 1 to the ROC space. We pose an example of a 3-label classification problem for simplicity. Let the label set $\mathcal{L}=\{\ell_1,\ell_2,\ell_3\}$ have generic labels where the indices 1, 2 and 3 do not necessarily imply any ordering. Assume that the sensor $\mathbf{s}$ and processor $\mathbf{p}$ mapped the disjoint events $\mathcal{E}_1,\mathcal{E}_2,\mathcal{E}_3$ into the feature set $\mathcal{F}=\mathbb{R}^2$ and produced sets that were not disjoint but distributed with the following mix of distributions:

Label	Centers (x,y)	Variance (x,y)	Density
ℓ_1	$(2,3)$	$(\frac{1}{4},1)$	$f_1(x,y)=\dfrac{\exp\left(-\left(\frac{x-2}{0.5}\right)\right)\exp\left(-\left(\frac{y-3}{1}\right)\right)}{(0.5)(1)\left[1+\exp\left(-\left(\frac{x-2}{0.5}\right)\right)\right]^2\left[1+\exp\left(-\left(\frac{y-3}{1}\right)\right)\right]^2}$
ℓ_2	$(-1,0)$	$(1,9)$	$f_2(x,y)=\frac{1}{2\pi(1)(3)}\exp\left(-\left(\frac{x+1}{1}\right)^2-\left(\frac{y-0}{3}\right)^2\right)$
ℓ_3	$(2,-2)$	$(1,1)$	$f_3(x,y)=\frac{1}{\pi^2(1)(1)}\left[1+\left(\frac{x-2}{1}\right)^2\right]\left[1+\left(\frac{y+2}{1}\right)^2\right]$

We seek values for $d_{1|1},d_{2|2}$ and $d_{3|3}$ such that

$$\begin{aligned}
&\frac{1}{2}d_{1|1}\left[\frac{2-\operatorname{erfc}(2\theta_1+2)}{2}\right]+\frac{1}{3}d_{2|2}\left[\frac{\operatorname{erf}(\theta_2)-\operatorname{erf}(\theta_1)}{2}\right]\\
&\quad+\frac{1}{6}d_{3|3}\left[\frac{1-\operatorname{erf}\left(\frac{\theta_2-1}{2}\right)}{2}\right]\\
&\quad+\frac{1}{12}\left[\operatorname{erf}\left(\frac{\theta_1-1}{2}\right)+2\operatorname{erf}(\theta_1)+6\operatorname{erfc}(2\theta_1+2)\right.\\
&\quad\left.+2\operatorname{erf}\left(\frac{\theta_2-1}{2}\right)-6\operatorname{erf}(\theta_2)-3\operatorname{erfc}(2\theta_2+2)+11\right]\\
&\qquad=1.
\end{aligned}$$

Multiplying by 12 and subtracting 12 from both sides yields

$$\begin{aligned}
&(6d_{1|1}+2d_{3|3}-1)+\operatorname{erf}\left(\frac{1}{2}\theta_1-\frac{1}{2}\right)+(2-2d_{2|2})\operatorname{erf}(\theta_1)\\
&\quad+(6-3d_{1|1})\operatorname{erfc}(2\theta_1+2)+(2-d_{3|3})\operatorname{erf}\left(\frac{1}{2}\theta_2-\frac{1}{2}\right)\\
&\quad-3\operatorname{erfc}(2\theta_2+2)+(2d_{2|2}-6)\operatorname{erf}(\theta_2)\\
&\qquad=0.
\end{aligned}$$

Since the error function, erf, the complemented error function, erfc, and the constant one function, 1, are linearly independent, we see that there are no values for $d_{1|1},d_{2|2}$ and $d_{3|3}$ for this equation to hold true. Therefore, there is no CC functional that will yield the same results as the Bayes performance functional for varying priors and misclassification costs. An optimal point using the CC functional cannot be found.

Examples 1 and 2 demonstrate that the Bayes cost performance functionals are more general than the CC functionals. This hinges on the use of the misclassifications versus the correct

The density distributions are graphed in Fig. 4. Define the classifier

$$\mathbf{a}_{\theta_1,\theta_2,\theta_3}(x,y)=\begin{cases}\ell_1 & \text{if } \theta_1\leq x,\quad y>\theta_2+\theta_3x,\\ \ell_2 & \text{if } x<\theta_1,\quad y\in\mathbb{R},\\ \ell_3 & \text{if } \theta_1\leq x,\quad y\leq\theta_2+\theta_3x.\end{cases}$$

This classifier creates a vertical plane that shifts horizontally with θ_1. It contains a ray that begins at (θ_1,θ_2) with slope θ_3. An

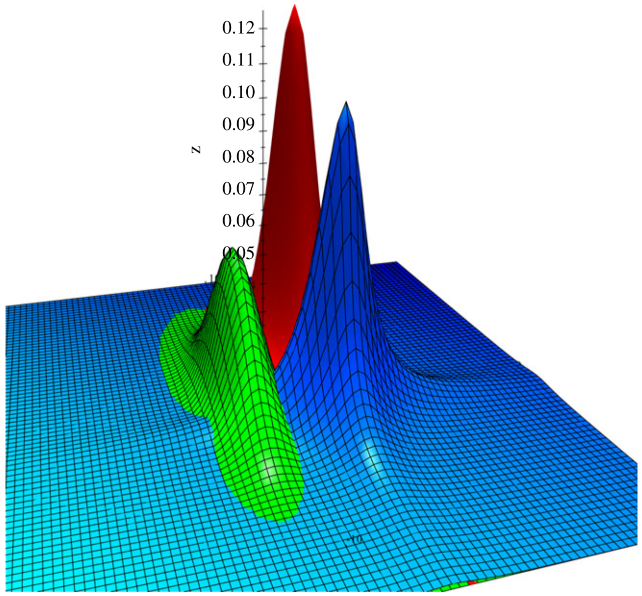

Fig. 4. The plots of the two-dimensional density functions: f_1 is Logistic, f_2 is Gaussian, and f_3 is Cauchy.

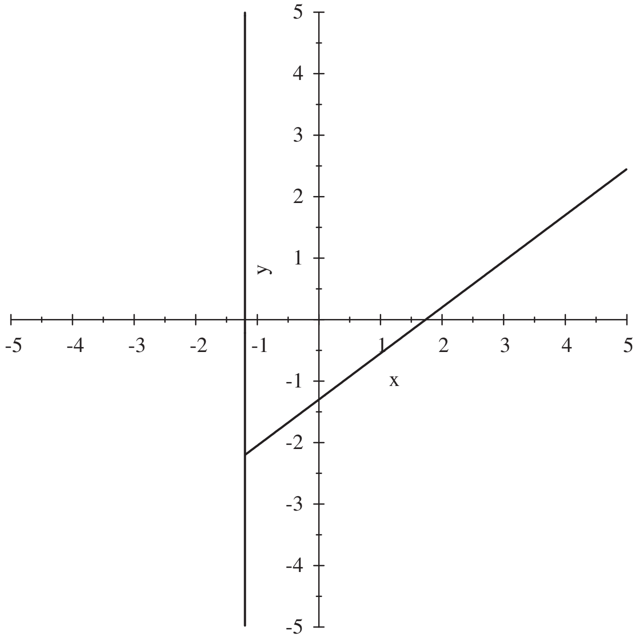

Fig. 5. The classifier $a_{\theta_1,\theta_2,\theta_3}$. The parameter θ_3 is the slope of the ray that begins at the point (θ_1,θ_2). The vertical line is located at $x=\theta_1$.

instantiation of this classifier is graphed in Fig. 5. Since the classifier has three classes then the dimension of the ROC space is $3(3-1)=6$. Since the classifier has three independent parameters it will produce a ROC manifold that has dimension 3, thus, its codimension is $6-3=3$.

Let $\boldsymbol{\theta}=(\theta_1,\theta_2,\theta_3)\in\Theta\equiv[-10,10]\times[-10,10]\times\mathbb{R}$, then define the pre-image

$$\mathbf{a}_{\boldsymbol{\theta}}^{\natural}(\ell_i)=\mathbf{a}_{\theta_1,\theta_2,\theta_3}^{\natural}(\ell_i)=\{(x,y)\in\mathbb{R}^2:\mathbf{a}_{\theta_1,\theta_2,\theta_3}(x,y)=\ell_i\}$$

and generally, the conditional probability as

$$P_{i|j}(\mathbf{A}_{\boldsymbol{\theta}})=\int_{\mathbf{a}_{\boldsymbol{\theta}}^{\natural}(\ell_i)}f_j(x,y)\,dy\,dx.$$

This generates the following six expressions for the misclassification probabilities:

$$P_{1|2}(\mathbf{A}_{\boldsymbol{\theta}})=\int_{\theta_1}^{\infty}\int_{\theta_2+\theta_3x}^{\infty}f_2(x,y)\,dy\,dx$$
$$=\int_{\theta_1}^{\infty}\frac{1}{2\sqrt{2\pi}}\exp\left(-\frac{(x+1)^2}{2}\right)\left[1-\operatorname{erf}\left(\frac{\theta_2+x\theta_3}{3\sqrt{2}}\right)\right]dx,$$

$$P_{1|3}(\mathbf{A}_{\boldsymbol{\theta}})=\int_{\theta_1}^{\infty}\int_{\theta_2+\theta_3x}^{\infty}f_3(x,y)\,dy\,dx$$
$$=\int_{\theta_1}^{\infty}\left[\frac{1}{\pi\left(1+\left(\frac{x-2}{1}\right)^2\right)}\right]\left[\frac{1}{2}-\frac{1}{\pi}\arctan(\theta_2+\theta_3x)\right]dx,$$

$$P_{2|1}(\mathbf{A}_{\boldsymbol{\theta}})=\int_{-\infty}^{\theta_1}\int_{-\infty}^{\infty}f_1(x,y)\,dy\,dx=\frac{1}{1+\exp\left(-\left(\frac{\theta_1-2}{0.5}\right)\right)},$$

$$P_{2|3}(\mathbf{A}_{\boldsymbol{\theta}})=\int_{-\infty}^{\theta_1}\int_{-\infty}^{\infty}f_3(x,y)\,dy\,dx=\frac{1}{\pi}\arctan(\theta_1-2)+\frac{1}{2}.$$

$$P_{3|1}(\mathbf{A}_{\boldsymbol{\theta}})=\int_{\theta_1}^{\infty}\int_{-\infty}^{\theta_2+\theta_3x}f_1(x,y)\,dy\,dx$$
$$=\int_{\theta_1}^{\infty}\frac{2\exp\left(-\left(\frac{x-2}{0.5}\right)\right)}{\left[1+\exp\left(-\left(\frac{x-2}{0.5}\right)\right)\right]^2}\left[\frac{1}{1+\exp\left(-\left(\frac{\theta_2+\theta_3x-3}{1}\right)\right)}\right]dx,$$

$$P_{3|2}(\mathbf{A}_{\boldsymbol{\theta}})=\int_{\theta_1}^{\infty}\int_{-\infty}^{\theta_2+\theta_3x}f_2(x,y)\,dy\,dx$$
$$=\int_{\theta_1}^{\infty}\frac{1}{2\sqrt{2\pi}}\exp\left(-\frac{(x+1)^2}{2}\right)\left[1+\operatorname{erf}\left(\frac{\theta_2+\theta_3x}{3\sqrt{2}}\right)\right]dx.$$

We consider the same Bayes performance functional as in Example 2 given by Eq. (10):

$$\Gamma=\begin{bmatrix}0 & c_{1|2} & c_{1|3}\\ c_{2|1} & 0 & c_{2|3}\\ c_{3|1} & c_{3|2} & 0\end{bmatrix}\begin{bmatrix}p_1 & 0 & 0\\ 0 & p_2 & 0\\ 0 & 0 & p_3\end{bmatrix}=\begin{bmatrix}0 & 1 & 3\\ 2 & 0 & 2\\ 1 & 3 & 0\end{bmatrix}\begin{bmatrix}\frac{1}{2} & 0 & 0\\ 0 & \frac{1}{3} & 0\\ 0 & 0 & \frac{1}{6}\end{bmatrix}$$
$$=\begin{bmatrix}0 & \frac{1}{3} & \frac{1}{2}\\ 1 & 0 & \frac{1}{3}\\ \frac{1}{2} & 1 & 0\end{bmatrix}.$$

Therefore, to find Bayes cost, we minimize $\langle\Gamma,\mathrm{R}(\mathbf{A}_{\boldsymbol{\theta}})\rangle$ over all $(\theta_1,\theta_2,\theta_3)\in\Theta$, that is, minimize

$$\langle\Gamma,\mathrm{R}(\mathbf{A}_{\boldsymbol{\theta}})\rangle=\operatorname{trace}\begin{bmatrix}0 & \frac{1}{3} & \frac{1}{2}\\ 1 & 0 & \frac{1}{3}\\ \frac{1}{2} & 1 & 0\end{bmatrix}^T\begin{bmatrix}0 & P_{1|2}(\mathbf{A}_{\boldsymbol{\theta}}) & P_{1|3}(\mathbf{A}_{\boldsymbol{\theta}})\\ P_{2|1}(\mathbf{A}_{\boldsymbol{\theta}}) & 0 & P_{2|3}(\mathbf{A}_{\boldsymbol{\theta}})\\ P_{3|1}(\mathbf{A}_{\boldsymbol{\theta}}) & P_{3|2}(\mathbf{A}_{\boldsymbol{\theta}}) & 0\end{bmatrix}$$
$$=\frac{1}{3}P_{1|2}(\mathbf{A}_{\boldsymbol{\theta}})+\frac{1}{2}P_{1|3}(\mathbf{A}_{\boldsymbol{\theta}})+P_{2|1}(\mathbf{A}_{\boldsymbol{\theta}})+\frac{1}{3}P_{2|3}(\mathbf{A}_{\boldsymbol{\theta}})$$
$$+\frac{1}{2}P_{3|1}(\mathbf{A}_{\boldsymbol{\theta}})+P_{3|2}(\mathbf{A}_{\boldsymbol{\theta}}).$$

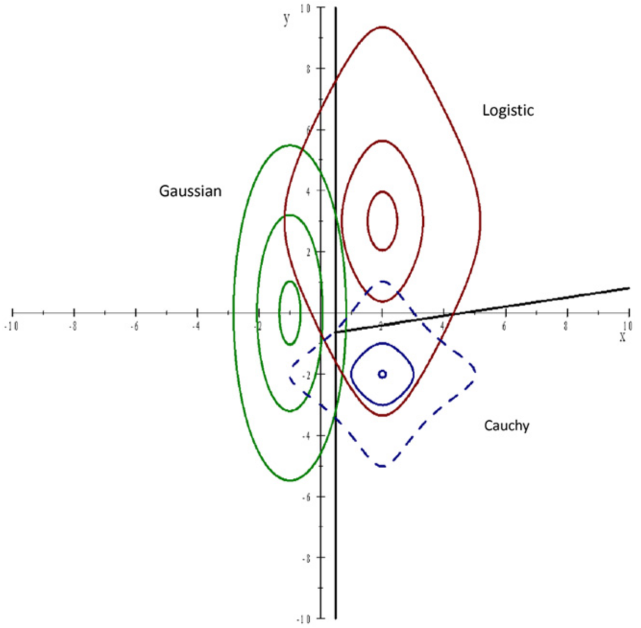

Fig. 6. The classifier $\mathbf{a}_{\theta_1,\theta_2,\theta_3}$ with the optimal choice occurring at $(\theta_1^*,\theta_2^*,\theta_3^*)=(0.48,-0.7,0.15)$. The contours of the distributions of the Logistic (solid), Gaussian (dash), and Cauchy (dots) occur at values 0.01, 0.05 and 0.1.

The resulting Bayes cost is

$$\varrho(\mathbb{A}) = \min_{\theta \in \Theta} \langle \Gamma, \mathrm{R}(\mathbf{A}_\theta) \rangle = 0.311$$

occurring at $(\theta_1^*, \theta_2^*, \theta_3^*) = (0.48, -0.7, 0.15)$. These optimal cut points are plotted in Fig. 6.

6. Data application: comparing classification systems for chronic allograft nephropathy

We apply the ROC manifold to data from medical diagnostics. Chronic allograft nephropathy (CAN) is a condition associated with late renal allograft loss. The ability to detect CAN non-invasively through gene expression levels in a set of urine sample markers would provide a more useful tool to assess renal injury as current use of serum creatinine levels provide only a late sign of renal injury. We investigate the detection of CAN in a subset of patients evaluated at least six months post-kidney transplant using several identified urine markers. In addition, we wish to detect the progression of CAN as measured through those individuals with proteinuria, a condition leading to CAN. Therefore, three outcome classes were identified: normal kidney function (NKF), NKF with proteinuria (NKF+), and CAN. Data collection and sample characteristics for 64 individuals have been described previously [22]. Truth data identified 32 individuals with NKF, 18 with NKF+ and 14 with CAN. Potential urine markers included in this analysis were angiotensingen (AGT) and transforming growth factor-beta (TGF-β).

The classification system for the kidney transplant patients can be described using the classification system theory in which the centrifuge acts as our sensor **s** mapping the urine sample for an individual into the raw data set $\mathcal{D}$ comprised mRNA preparations. These preparations are not refined enough for our classification, thus, the processor **p** is used to conduct real-time polymerase chain reaction (PCR) analysis in order to map the data from $\mathcal{D}$ into a feature set $\mathcal{F}$ comprised possible analytes, such as AGT and TGF-β. These features are subjected to a classifier $\mathbf{c}_\theta$ where $\theta \in \Theta$ and the parameter set Θ represents the set of possible threshold values. Finally, the classifier, $\mathbf{c}_\theta$, outputs a label from the label set $\mathcal{L} = \{\text{NKF}, \text{NKF}+, \text{CAN}\}$. We denote the classification system using AGT as system **A**, and the classification system using TGF-β as system **B**.

To demonstrate the use of the ROC manifold to find optimal thresholds for classification, a simple rectangular classifier was chosen so that the resulting threshold values would represent actual marker values to distinguish between the three groups. First, AGT and TGF-β were considered separately. To distinguish between the three labels for this simple case, the rectangular classifier was applied separately to each biomarker, Y, such that if $Y \le \theta_1$ then we assigned the label NKF, if $\theta_1 < Y \le \theta_2$ we assigned the label NKF+, and if $Y > \theta_2$ we assigned the label CAN. This process was repeated for every $\theta_1, \theta_2 \in \Theta = (-\infty, \infty)$ with $\theta_1 < \theta_2$. Theoretically, θ_1 and θ_2 can range from $(-\infty, \infty)$, however, no new information regarding the classification of the observations is obtained outside the range of Y, thus, θ_1 and θ_2 can be restricted to the observable range of Y, in this case $\theta_1, \theta_2 \in [0, \max(Y)]$. Once we classified each observation for every θ_1, θ_2 pair, we computed the resulting misclassifications (for three labels we had $3(3-1) = 6$ errors). These six-tuples comprised the points making up the ROC manifold. Applying our Bayes cost performance functional to the ROC manifold, we found optimal threshold values for AGT and TGF-β. These threshold values denote cut points of the biomarker that minimize the misclassification error in CAN progression (NKF, NKF+, CAN) according to our performance functional. Using the resulting optimal threshold values, we assessed the correct classification rate and the misclassification rate for each label. Results for the Bayes cost functional, assuming equal costs and priors among the three classes, are given in Table 1.

The optimal threshold values for AGT identify the following for classification: if $\text{AGT} \le 3.3$ then we assign the label NKF, if $3.3 < \text{AGT} \le 55.8$ we assign the label NKF+, and if $\text{AGT} > 55.8$ we assign the label CAN. For TGF-β, the optimal threshold values identify the following classifications: if TGF-$\beta \le 1.2$ then we assign the label NKF, if $1.2 < \text{TGF}-\beta \le 2.5$ then we assign the label NKF+, and if $\text{TGF}-\beta > 2.5$ then we assign the label CAN.

Whether we use the ROC manifold or the CC manifold, we will find the same optimal threshold values for these markers under the assumption of equal costs and priors. Therefore, the correct classification and misclassification rates at this optimal threshold will be the same for the ROC manifold and the CC manifold.

When we ignore the error classification and observe only the correct classification rates (Table 1), we conclude that each marker performs slightly better than chance in the overall classification of these patients. Invoking our Bayes cost performance functional as our measure of performance, we determine that TGF-β is a better classifier than AGT because

$$\varrho(\mathbb{B}) = 6.32 < \varrho(\mathbb{A}) = 6.96.$$

Recall that we assumed the costs for misclassification and prior probabilities across the classes would be equal. Although AGT classifies more than twice as many NKF patients correctly as TGF-β, we observe that AGT misclassifies every patient with CAN (100% misclassification) whereas TGF-β classifies all but three CAN patients correctly (21.4% misclassification). The minimum risk assuming equal errors and costs across classes then identifies TGF-β as a better marker.

Next, we examined the use of both markers to separate patients into these three classes in hopes of minimizing Bayes cost further. Using an analogous classifier as that for the single marker, but now extended to two markers, we use both AGT (x) and TGF-β (y) to classify these patients. Thus, four parameters are now identified, θ_1 and θ_2 for AGT and ϕ_1 and ϕ_2 for TGF-β. We denote this classification system as system **C**, and propose the following form for this classifier, $\mathbf{c}_{\theta_1,\theta_2,\phi_1,\phi_2}(x,y)$:

$$\mathbf{c}_{\theta_1,\theta_2,\phi_1,\phi_2}(x,y) = \begin{cases} \text{NKF} & \text{for } -\infty < x < \infty,\ \phi_2 < y < \infty, \\ \text{NKF} & \text{for } \theta_1 \le x < \theta_2,\ \phi_1 \le y < \phi_2, \\ \text{NKF}+ & \text{for } -\infty < x < \theta_1,\ -\infty < y < \phi_2, \\ \text{CAN} & \text{for } \theta_1 < x < \infty,\ -\infty < y < \phi_1, \\ \text{CAN} & \text{for } \theta_2 < x < \infty,\ -\infty < y < \phi_2. \end{cases}$$

Table 1
Correct classification and misclassification rates by class for single markers.

Classification system	Bayes cost	Parameters		Correct classification rate	Misclassification rate within class		
		θ_1	θ_2		NKF ($n=32$)	NKF+ ($n=18$)	CAN ($n=14$)
A (AGT)	6.96	3.3	55.8	62.5 ($n=40$)	9.4 ($n=3$)	38.9 ($n=7$)	100.00 ($n=14$)
B (TGF-β)	6.32	1.2	2.5	54.7 ($n=35$)	59.4 ($n=19$)	38.9 ($n=7$)	21.4 ($n=3$)

Table 2
ROC manifold classification rates (%) for combined thresholds of AGT and TGF-β.

Costs and priors	Correct classification	Misclassifications within class			Parameters			
		NKF	NKF+	CAN	θ_1	θ_2	ϕ_1	ϕ_2
Equal costs	68.8 (n=44)	9.4 (n=3)	61.1 (n=11)	43.0 (n=6)	2.0	2.6	4.3	4.0
Unequal costs	57.8 (n=37)	56.3 (n=18)	50.0 (n=9)	0.00 (n=0)	1.2	1.3	9.5	10.5

All combinations of these four parameters were examined over their respective ranges. The six misclassification rates for every combination of the four parameters were determined. Bayes cost was computed first assuming equal costs and priors among the six errors and then with differential costs of misclassification.

Results using both markers simultaneously are given in Table 2. Under the assumption of equal costs and priors, AGT and TGF-β together classify 68.8 (n=44) of the patients correctly and misclassify 20 patients. This is an improvement in the total correct classification rate for equal costs and priors over the use of any single marker (Table 1). Using AGT and TGF-β together, misclassification rates improve within class differentially, with fewer patients being misclassified as NKF as compared to TGF-β, and fewer patients being misclassified as CAN as compared to AGT. Most importantly, we compare the Bayes cost for this classification system to the two prior systems in which single markers were used. Bayes cost for classification system $\mathbb{C}$ is 6.04. Thus,

$$\varrho(\mathbb{C}) < \varrho(\mathbb{B}) < \varrho(\mathbb{A})$$

and we conclude that the classification system using both AGF and TGF-β is better than either proposed system using these markers separately.

As a separate study, we may be concerned about high class-specific misclassification rates, as we would rather not misclassify patients who truly have CAN or who are progressing towards CAN (i.e., those NKF+). This suggests a different cost structure for the errors associated with these two classes over the common equal cost assumption.

Therefore, we imposed a cost structure to the classes. We assumed the cost for making the errors of classifying CAN as either NKF or NKF+ was five times worse than NKF and that making the errors of classifying NKF+ as CAN or NKF as two times worse. Applying these costs to our six-tuples for each set of parameters, another optimal set of threshold values was found (Table 2). For this set, we observe fewer misclassifications of NKF+ and no patient with CAN is misdiagnosed.

This weighted combined classifier resulted in the following decision thresholds:

$$\mathbf{c}_{\theta_1,\theta_2,\phi_1,\phi_2}(x,y) = \begin{cases} \text{NKF} & \text{for } -\infty < x < \infty,\ 10.5 < y < \infty, \\ \text{NKF} & \text{for } 1.2 \le x < 1.3,\ 9.5 \le y < 10.5, \\ \text{NKF+} & \text{for } -\infty < x < 1.2,\ -\infty < y < 10.5, \\ \text{CAN} & \text{for } 1.2 < x < \infty,\ -\infty < y < 9.5, \\ \text{CAN} & \text{for } 1.3 < x < \infty,\ -\infty < y < 10.5. \end{cases}$$

The classifiers used in this example do not necessarily identify the optimal classifier for this problem, though they illustrate several concepts related to quantifying the performance of classification systems. Firstly, the computation of Bayes cost is straightforward and provides a way to compete classification systems when visual inspection methods do not exist due to the dimensionality of the system. Further, Bayes cost offers a way to compare the performance of the system for the user's underlying assumptions regarding costs and class-specific prevalence. Secondly, the use of the ROC manifold is a better method for finding optimal parameters when using unequal costs of misclassification and prior probabilities. The CC method cannot be used to find such values that minimize these important misclassifications as it cannot consider varying the costs and priors associated with each of the possible n^2-n errors. Lastly, it is worth noting that the types of classifiers compared in this example are useful in a clinical setting as the optimal threshold parameters identified relate directly to the observable levels of the biomarkers within the patient. The research for using these biomarkers to classify patients as to their CAN progression is not completed. Future work for this application includes competing other classifiers, possibly including decision tree and modelling methods, considering several other biomarkers along with AGT and TGF-β to improve the classification of patients with these conditions, and refining appropriate cost and prior probability assumptions.

7. Conclusions

The ROC manifold offers several advantages over the CC manifold as a tool for classification system performance due to the fact that the ROC manifold is an object in $\mathbb{R}^{n^2-n}$, isomorphic to the hypercube $[0,1]^{n^2-n}$ we call ROC space. The ROC manifold does not project down in dimension from ROC space, unlike the CC manifold, which is a factor of $n-1$ less in dimension. The ROC manifold maintaining this dimensionality enables the researcher to use it to consider varying costs within a class and does not require those costs to be constant. Presumably, one can identify more accurate operating parameters when the costs reflect the real-world situation, and there may be many cases where such weighting is warranted. Optimal points on the CC manifold are only tractable when costs and prior probabilities are equal. In this case, computing these points on the projected manifold in n-dimensional space is computationally easier (considering computational time and power) than computing these points on the ROC manifold in (n^2-n)-dimensional ROC space. From the ROC manifold, the CC manifold can be created using the conjunctive equations comprising the classification system. No information is lost as to the relative classification of the system in relation to each class when using the ROC manifold. This is not true for the CC manifold.

The CC manifold can be used to find markers that perform well over all threshold values, via comparison of the VUS. VUS has no direct meaning for the ROC manifold [9]. However, VUS for the CC manifold is not always obtainable depending on the dimensionality of the classification system. For example, a classification system with one parameter and three classes has a CC manifold (topologically a 1-manifold) which is a trajectory in 3-space and clearly has no volume. VUS for this system is zero, though applying other performance functionals to this trajectory, such as Bayes cost, is possible. Further, when comparing classification systems, unless there is system dominance with respect to VUS, that is, one system has a CC manifold with better correct classification rates across all combinations of parameters, it is possible that the system with the smaller VUS may have better performance for specific ranges of operating parameters [15]. If these operating parameters are of interest, then use of the CC manifold VUS to choose the best classification system may be misleading. For this reason we describe and compute Bayes cost as a more useful performance functional with which to judge the classification system in terms of the best performance that the system can deliver. We compare and compete systems based on the system performance at the optimal point(s).

We have perhaps belabored the comparisons between the CC manifold and the ROC manifold, specifically through the use of VUS, despite earlier work [9] and ourselves demonstrating flaws with this measure. However, some disciplines (statistics) continue to focus and publish on the use of the CC manifold and computations of VUS despite these flaws. Therefore we have chosen to describe the relationship between these manifolds and the dual problem in order to promote understanding of these methodological tools and what these tools can and cannot tell us in terms of classification system performance. We offer the ROC manifold as a method for determining classification system performance for even more general systems in which likelihood criteria may not be available or the manifold has higher codimensionality with the ROC space, e.g. through the number of parameters, extending previous work in ROC manifolds.

The examples in this paper were chosen to demonstrate the methods of using and computing the ROC manifold and associated performance functionals with which to compete systems. We offer examples in which likelihood criteria are readily traceable, in which codimensionality is greater than 1, and for which we can compute optimal system performance using both equal and unequal costs and prior probabilities. The classifiers chosen here are not meant to be suggested as optimal classification systems (we know they are not). Indeed our intent is not to offer a method in which to identify the optimal classification system, but rather to be able to describe and compare classification system performance for specific systems of interest within the family of systems. As such, the optimal parameters we find through Bayes cost allow us to identify an optimal setting for each system of interest. Work is ongoing for the CAN data example to find better classifiers using the ROC manifold and associated performance functionals with which to compare candidate classifiers.

Finally, the methods outlined in this paper allow us to find optimal parameters for a classification system with any finite number of features and labels, given a particular functional with which to measure the performance of the CSF. The advantages include not having to examine one parameter at a time, or having to assume a specific relationship among the parameters. The methods leverage the power of looking through all of the parameters simultaneously provided all of the class conditional probabilities are computed, which is only limited by computing power and time, depending on the number of classes proposed. Future work will extend the ROC manifold to combined classification systems (not just combined features of systems) in order to determine how the performances of combined classification systems compare to single classification systems and under what conditions combined systems may outperform that of the single systems.

References

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

[2] C.T. Nakas, C.T. Yiannoutsos, Ordered multiple-class ROC analysis with continuous measurements, Statistics in Medicine 23 (2004) 3437–3449.

[3] C. Xiong, G. van Belle, J.P. Miller, J.C. Morris, Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups, Statistics in Medicine 25 (2006) 1251–1273.

[4] P.S. Heckerling, Parametric three-way receiver operating characteristic surface analysis using mathematica, Medical Decision Making 21 (2001) 409–417.

[5] X. He, C.E. Metz, B.M.W. Tsui, J.M. Links, E.C. Frey, Three-class ROC analysis—a decision theoretic approach under the ideal observer framework, IEEE Transactions on Medical Imaging 25 (5) (2006) 571–581.

[6] S. Dreiseitl, L. Ohno-Machado, M. Binder, Comparing three-class diagnostic tests by three-way ROC analysis, Medical Decision Making 20 (2000) 323–331.

[7] C.T. Nakas, T.A. Alonzo, ROC graphs for assessing the ability of a diagnostic marker to detect three disease classes with an umbrella ordering, Biometrics 63 (2007) 603–609.

[8] D.C. Edwards, C.E. Metz, M.A. Kupinski, Ideal observers and optimal ROC hypersurfaces in N-class classification, IEEE Transactions on Medical Imaging 23 (7) (2004) 891–895.

[9] D.C. Edwards, C.E. Metz, R.M. Nishikawa, The hypervolume under the ROC hypersurface of "near-guessing" and "near-perfect" observers in N-class classification tasks, IEEE Transactions on Medical Imaging 24 (3) (2005) 293–299.

[10] B.K. Scurfield, Multiple-event forced-choice tasks in the theory of signal detectability, Journal of Mathematical Psychology 40 (3) (1996) 253–269.

[11] S.N. Thorsen, M.E. Oxley, A description of competing fusion systems, Information Fusion 7 (2006) 346–360.

[12] X. He, E.C. Frey, The meaning and use of the volume under a three-class ROC surface (VUS), IEEE Transactions on Medical Imaging 27 (5) (2008) 577–588.

[13] T. Landgrebe, R. Duin, A simplified volume under the ROC hypersurface, SAIEE Africa Research Journal 98 (3) (2007) 94–100.

[14] C. Ferri, J. Hernández-Orallo, M.A. Salido, Volume under the ROC surface for multi-class problems. Exact computation and evaluation of approximations, Technical Report, Dep. Sistemes Informatics i Computacio, Univ. Politecnica de Valencia, Spain, 2003.

[15] D.K. McClish, ROC volumes—Should they be used? Biostatistics 49 (5) (2007) 665–666.

[16] S.I. Resnick, A Probability Path, Birkhauser, Boston, MA, 2005.

[17] S.N. Thorsen, M.E. Oxley, A description of competing fusion systems, Information Fusion Journal 7 (2006) 346–360.

[18] J.P. Egan, Signal Detection Theory and ROC Analysis, Academic Press, New York, 1975.

[19] D.G. Luenberger, Optimization by Vector Space Methods, John Wiley and Sons, Inc., New York, 1969.

[20] H. Minkowski, Allgemeine lehrsätze über konvexe polyeder, Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen (1897) 198–219.

[21] J.R. Munkres, Topology. A First Course, Prentice-Hall, Inc., Englewood Cliffs, 1975.

[22] V.R. Mas, L.A. Mas, K.J. Archer, K. Yanek, A.L. King, E.M. Gibney, A. Cotterell, R.A. Fisher, M. Posner, D.G. Maluf, Evaluation of gene panel mRNAs in urine samples of kidney transplant recipients as a non-invasive tool of graft function, Molecular Medicine 13 (5–6) (2007) 315–324.

Christine M. Schubert earned her Ph.D. in Applied Mathematics from the Air Force Institute of Technology and is currently an assistant professor in the Department of Mathematics and Statistics at the Air Force Institute of Technology. Her research interests and background are in the measurement of performance for classification systems and methods to improve classification.

Steven N. Thorsen earned his Ph.D. in Applied Mathematics from the Air Force Institute of Technology and is currently an assistant professor in the Department of Mathematical Sciences at the US Air Force Academy in Colorado Springs, Colorado. His research interests and background are in the optimization of classification systems and the theory of information fusion.

Mark E. Oxley earned his Ph.D. in Mathematics from North Carolina State University and currently is a Professor of Mathematics in the Department of Mathematics and Statistics at the Air Force Institute of Technology. His research interests are in information fusion, ROC analysis, functional analysis, and partial differential equations.