AFIT/GAM/ENC/08-01

RISK-BASED COMPARISON OF CLASSIFICATION SYSTEMS

THESIS

Seth B. Wagenman, Second Lieutenant, USAF
AFIT/GAM/ENC/08-01

DEPARTMENT OF THE AIR FORCE

AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government.

AFIT/GAM/ENC/08-01

RISK-BASED COMPARISON OF CLASSIFICATION SYSTEMS

THESIS

Presented to the Faculty
Department of Mathematics and Statistics
Graduate School of Engineering and Management
Air Force Institute of Technology
Air University
Air Education and Training Command
In Partial Fulfillment of the Requirements for the
Degree of Master of Science

Seth B. Wagenman, B.S.
Second Lieutenant, USAF

March 2008

AFIT/GAM/ENC/08-01

RISK-BASED COMPARISON OF CLASSIFICATION SYSTEMS

Seth B. Wagenman, B.S.
Second Lieutenant, USAF

Approved:

Steven N. Thorsen (Chair)   Date

Mark E. Oxley (Member)   Date

David M. Kaziska (Member)   Date

AFIT/GAM/ENC/08-01

Abstract

Performance measures for families of classification system families that rely upon the analysis of receiver operating characteristics (ROCs), such as area under the ROC curve (AUC), often fail to fully address the issue of risk, especially for classification systems involving more than two classes. For the general case, we denote matrices of class prevalences, costs, and class-conditional probabilities, and assume costs are subjectively fixed, acceptable estimates for expected values of class-conditional probabilities exist, and mutual independence between a variable in one such matrix and those of any other matrix. The ROC Risk Functional (RRF), valid for any finite number of classes, has an associated parameter argument, that which specifies a member of a family of classification systems, and which system minimizes Bayes risk over the family. We typify joint distributions for class prevalences over standard simplices by means of uniform and beta distributions, and create a family of classification systems using actual data, testing independence assumptions under two such class prevalence distributions. We minimize risk under two different sets of costs.

Acknowledgments

I would like to sincerely thank Maj Steve Thorsen for his patience while guiding this research, as well as Dr. Mark Oxley and Maj Dave Kaziska for their help in presenting it. I would also like to thank Dr. Ken Bauer and Maj Sam Wright for always being there to aid in reality-checking my implementation of these concepts, and Dr. Aihua Wood and Dr. Matt Fickus for helping me to know that the Theorems in the Appendix can be proved for all positive integers, for providing me with the confidence and ability to prove one of them fully, and for comforting me with the thought that someone else has probably already proved the other one for integers greater than 47.

Most importantly, I thank God for helping me to see things I never could have without His help, and for giving me my dear wife, who listened to my incessant babbling about this thesis and endured many a lonely evening whilst I fought with my computer, and who generally motivated, inspired, encouraged, and enabled me to seek education at the graduate level.

A note of thanks is also in order to the Air Force Office of Scientific Research for supporting our research efforts.

Seth B. Wagenman

Table of Contents

List of Figures

Figure Page

List of Tables

List of Notations or Symbols

Notation or Symbol		Meaning
1.	AER	Actual Error Rate
2.	AUC	Area Under the ROC Curve
3.	FN	False Negative Count
4.	*fnr*	False Negative Rate
5.	FP	False Positive Count
6.	*fpr*	False Positive Rate
7.	$\langle\, ,\, \rangle_F$	Frobenius Dot Product
8.	$\odot$	Hadamard Matrix Product
9.	PCA	Principal Components Analysis
10.	PNN	Probabilistic Neural Net
11.	ROC	Receiver Operating Characteristic
12.	ROCCH	ROC Convex Hull
13.	RRF	ROC Risk Functional
14.	Δ_{m}	Standard m-Simplex
15.	TN	True Negative Count
16.	*tnr*	True Negative Rate
17.	TP	True Positive Count
18.	*tpr*	True Positive Rate
19.	VUS	Volume Under the Surface

RISK-BASED COMPARISON OF CLASSIFICATION SYSTEMS

I. Introduction and Mathematical Foundations

The concept of risk is a major feature of Bayesian decision theory [5, pp. 24-28], [18, p. 437]. Its power is evident in that it takes into account not only the relative severity of expected conditional losses for each possible decision, but also the likelihood of events upon which the occurrence of each loss is conditioned. It allows definition of these quantities through the use of either discrete or continuous random variables, or a combination of both. In this way, it accounts for many *characteristics* of the *operating* environment.

The term Receiver Operating Characteristic (ROC) seems to refer directly to this type of decision-theoretical framework, yet practical applications of decision theory in which this term appears often ignore critical aspects of Bayesian theory. To show this, a brief introduction to ROC analysis is necessary, as is a precise set of mathematical definitions, to establish a framework for possible correction of these oversights.

1.1 Introduction

The field of Receiver Operating Characteristic analysis emerged in the 1940s, during early attempts to discern between the presence or absence of signals amidst noise [6, pp. 1-2]. Since there are only two possible outcomes, such a signal detection process is an example of two-class or *binary* classification.

In signal detection, there are two possible classification *errors*—falsely perceiving the presence of signals amidst noise *when there is only noise*, and failing to detect a signal in the midst of noise. One representation of the ROC for a binary classification system is simply a vector of the likelihoods of these errors. If the method of classification changes, so may these likelihoods, thereby generating a different ROC vector. A collection of estimates for such ROC vectors plotted on a unit square may offer limited visual insight into comparison of the classification methods whose *characteristic* behavior they represent, and even more when other factors of interest are plotted on a third axis [10]. Many authors have developed advanced geometrical frameworks relating to the points so plotted, due to the common practice of calculating areas under a curve constructed of these plotted points [8], [10], [12], [13], [14], [21], [22], [26], [34], [36]. The use of ROCs in such comparative decision-making is referred to as ROC analysis.

Even though ROC analysis is used in many fields to compete binary classification techniques, it appears that very few of its proponents have fully realized the importance and potential of risk-based comparison as a tool for comparing classification techniques, especially those in which there are more than two distinct classes [30, pp. 57-64], [31, p. 352], [32, p. 4]. Although practical risk-based comparison of classification systems requires what could be considered unrealistic independence assumptions to enable the risk calculations, the possible insights gained when these assumptions are met may at least justify the expense of testing them, and when viewed in light of the implicit assumptions connected with a failure to fully consider all elements of the risk calculation, these assumptions may not be harsh at all. Since the failure to meet these assumptions is rarely mentioned in modern ROC analysis literature, the reason for not calculating risk may simply be the lack a practical and rigorous mathematical framework for its analysis. There is recent work, however, which constitutes a foundation on which to build a framework for measuring the performance of classification systems, with Bayes risk as the measure of optimality, and incorporating some of the independence assumptions mentioned above [32]. The intent of this thesis is neither to show how these assumptions may be met, nor to stipulate as to the relative importance of actually meeting them, but instead to show how they may be tested, and then to assume that even if they are not met, the disadvantage of such failure is mitigated by the ability to easily calculate risk. The major point of interest is that geometric analyses are replaced by risk-based comparisons, thereby possibly lessening the need to construct curves or surfaces, or to calculate geometric quantities.

1.2 Definitions and Assumptions

Before proceeding, it is necessary to define notation and terms relating to general classification theory and ROC analysis. Examples from the field of signal detection will illustrate selected concepts.

Definition 1 (Experiment)**.** An experiment is a complex of reproducible conditions resulting in a set of well-defined outcomes [16, pp. 3-5], [29, p. 32].

For example, the presence of electromagnetic radiation constitutes a possible signal detection experiment.

Definition 2 (Elementary Event)**.** An elementary event is an experimental outcome which cannot be further decomposed into other, more basic experimental outcomes [33, p. 26].

An elementary event in signal detection could be $e =$ *a detectable instance of electromagnetic radiation exists.*

Definition 3 (Event Set)**.** An event set is the set $\mathrm{E} = \{e_\lambda\}_{\lambda \in \Lambda}$ of possible elementary events resulting from a given experiment, where the index set Λ may be uncountable [16, pp. 3-5], [29, p. 33], [33, p. 27].

Definition 4 (Sensor, Data Set)**.** A function s with event set domain E, whose action is to observe elementary events e and gather data about them; therefore, the range of a sensor is a set D of data elements d_e corresponding to elementary events observed [32, p. 1].

In signal detection, a data set could be a hard disk containing information gathered through a cable connected to a radio signal detection machine.

Definition 5 (Processor, Feature Set)**.** Given a data set D, a processor is a function p with data subset domain $\mathrm{D}' \subseteq \mathrm{D}$, whose action is to transform data corresponding to elementary events e observed by a sensor $s\colon \mathrm{E} \longrightarrow \mathrm{D}$ (whose range is an event set E and whose range contains the domain of p) into a vector of numbers, usually real-valued; therefore, the range of a processor is a feature set F of finite-dimensional vectors f_e corresponding to elementary events e whose representational data points d_e are elements of the domain of p [32, p. 1]. An element $f_e \in \mathrm{F}$ of a feature set is called an *exemplar.*

In signal detection, a processor could be a computer that receives a floppy disk containing some of the data gathered by a sensor and performs calculations to produce a matrix of real numbers, with columns corresponding to wave amplitude and frequency variables, and with exemplars as row vectors in the matrix corresponding to elementary events observed by the sensor. Part of these calculations may also create new variables that are related to, but not defined strictly the same as, the variables observed by the sensor. For example, Principal Components Analysis (PCA) is a method of reducing the number of features, by creating a few linear combinations of them which explain most of the variance in the original features matrix. The coefficients of each of these linear combinations are applied to each row of the original feature data to produce a principal component score, which in turn becomes a new feature variable.

Definition 6 (Event)**.** Any subset $\mathcal{A} \subset \mathrm{E}$ of an event set is called an event [29, p. 34]. Note that *any* set E is always regarded as a subset $\mathrm{E} \subset \mathrm{E}$ of itself, and the empty set $\varnothing$ is a subset of every set besides itself, even though we may not explicitly denote its presence. Also, for $\mathcal{A} \subset \mathrm{E}$, if *any* elementary event $e \in \mathcal{A}$ occurs, then $\mathcal{A}$ has also occurred.

In signal detection, the sets $\mathcal{E}_1 = \{\textit{radiation with signals amidst noise is present}\}$, and $\mathcal{E}_2 = \{\textit{only noise is present}\}$ are all subsets of the event set $\mathrm{E} = \mathcal{E}_1 \cup \mathcal{E}_2$, as is the set E itself; therefore, *each* set listed above constitutes *an event.*

Definition 7 (Finite Set Partition)**.** Given a non-empty set E and a finite index set Λ, a collection of subsets $\big\{\mathcal{E}_\lambda \subset \mathrm{E}\big\}_{\lambda\in\Lambda}$ is a finite set partition of E when the following hold:

(i) $\mathcal{E}_\lambda \cap \mathcal{E}_\mu = \varnothing, \quad \forall\ \mu, \lambda \in \Lambda \ \ni\ \mu \neq \lambda$, and

(ii) $\displaystyle\bigcup_{\lambda\in\Lambda} \mathcal{E}_\lambda = \mathrm{E}$

i.e., $\big\{\mathcal{E}_\lambda \subset \mathrm{E}\big\}_{\lambda\in\Lambda}$ is a finite collection of mutually exclusive subsets of E whose union is the whole set E [29, p. 36].

Definition 8 (Classifier, Label Set)**.** A classifier is a function c with feature subset domain $\mathrm{F}' \subset \mathrm{F}$, whose action assigns exactly one label ℓ out of a *finite* set L of *distinct* labels to each feature vector $f \in \mathrm{F}'$; therefore, a label set $\mathrm{L} = \{\ell_1, \ell_2, \ell_3, \ldots, \ell_\mathrm{n}\}$ is the range of an n-class classifier $c\colon \mathrm{F}' \longrightarrow \mathrm{L}$, such that if an event set E is the domain of a sensor $s\colon \mathrm{E} \longrightarrow \mathrm{D}$ whose range contains the domain $\mathrm{D}' \subset \mathrm{D}$ of a processor $p\colon \mathrm{D}' \longrightarrow \mathrm{F}$, whose range in turn contains the domain $\mathrm{F}' \subset \mathrm{F}$ of c, then L partitions E into a set of n mutually exclusive subsets $\big\{\mathcal{E}_\mathrm{j}\big\}_{\mathrm{j}=1}^{\mathrm{n}}$ called *classes*, whose union is the entire event set E, such that each class $\mathcal{E}_{\mathrm{j}} \subset \mathrm{E}$ corresponds to exactly one label $\ell_{\mathrm{j}} \in \mathrm{L}$ [32, pp. 1-2].

A signal detection classifier could be an artificial neural network operating on rows (exemplars) extracted from a principal component score matrix whose row vectors correspond to particular *instances* of electromagnetic radiation. Note that the method of creating or *training* such a classifier, as well as testing it against a subset of the data from which it is created, is subjective; for example, a binary classifier could be flipping a fair coin.

The signal detection label set $\mathrm{L} = \{\ell_1, \ell_2\}$ (where elementary events in class $\mathcal{E}_1 = \{\textit{radiation with signals amidst noise is present}\}$ correspond to the label ℓ_1, and elementary events in class $\mathcal{E}_2 = \{\textit{only noise is present}\}$ correspond to the label ℓ_2) induces the finite set partition $\mathrm{E} = \mathcal{E}_1 \cup \mathcal{E}_2$, where the event set is $\mathrm{E} = \{\textit{electromagnetic radiation is present}\}$. This is an example of a *two-class* partition.

Definition 9 (Classification System)**.** Given the following:

(i) a sensor $s\colon \mathrm{E} \longrightarrow \mathrm{D}$ with event set domain E and data set range D,

(ii) a processor $p\colon \mathrm{D}' \longrightarrow \mathrm{F}$ with data subset domain $\mathrm{D}' \subset \mathrm{D}$ and feature set range F, and

(iii) a classifier $c\colon \mathrm{F}' \longrightarrow \mathrm{L}$, with feature subset domain $\mathrm{F}' \subset \mathrm{F}$ and label set range L,

the composition $A = c \circ p \circ s : \mathrm{E} \longrightarrow \mathrm{L}$ with event set domain E and label set range L is a classification system $A\colon \mathrm{E} \to \mathrm{L}$ [32, p. 2].

Definition 10 (Threshold Set)**.** Given any feature set F, a threshold set Θ of interest is a set of parameters $\theta \in \Theta$ that influence mappings with domain F. These parameters need be neither univariate, continuous, nor real-valued.

A signal detection threshold parameter θ_1 that is *neither* continuous *nor* real-valued could be choosing whether to flip a quarter or a nickel, whereas a continuous *and* real-valued parameter θ_2 might be the choice of a real-valued discriminating criterion [5, pp. 48-49]. Some types of artificial neural net classifiers have a continuous parameter called the *spread*, such that each setting of this parameter effectively defines a new classifier, given a particular choice of methods for training. A threshold set Θ of interest might also be the Cartesian product:

$$\Theta = \Big\{ \big(\theta_1, \theta_2\big) : \theta_1 \in \Theta_1, \, \theta_2 \in \Theta_2 \Big\} = \Theta_1 \times \Theta_2$$

of threshold sets Θ_1 and Θ_2 [32, pp. 1-2]. It should be noted that in practice, we only consider a finite number of threshold parameters to approximate a continuous threshold set, and so each distinct finite sample may be considered a separate discrete threshold set.

Definition 11 (Family of Classification Systems Over a Threshold Set)**.** Given a threshold set Θ such that the value of the parameter $\theta \in \Theta$ determines the action of a classifier c_θ, a family of classification systems of the form $A_\theta \equiv c_\theta \circ p \circ s$ over the threshold set Θ is the collection $\mathbb{A}_\Theta = \big\{ A_\theta \colon \theta \in \Theta \big\}$ of all such classification systems.

It should be noted here that when searching for a classification system A_θ to meet some particular criterion from an infinite family A_Θ defined over a continuous threshold set Θ, practicality requires the creation of finite families of classification systems over discrete samples from the continuous threshold set. Even though these samples are subsets of the same set, they may be distinct, and thus the families of classification systems over these sample threshold sets are also distinct.

Definition 12 (σ-Field)**.** Given a non-empty set E and a *countable* index set Λ, a collection $\mathscr{E}$ of subsets $\mathcal{A} \subset \mathrm{E}$ is a σ-field over E when the following hold true:

(i) $\mathrm{E} \in \mathscr{E}$,

(ii) if $\mathcal{A} \in \mathscr{E}$, then $\mathcal{A}^{\mathrm{C}} \in \mathscr{E}$, and

(iii) $\mathcal{A}_\lambda \in \mathscr{E}\ ,\ \forall\ \lambda \in \Lambda \quad \Longrightarrow \quad \bigcup_{\lambda \in \Lambda} \mathcal{A}_\lambda\ \in\ \mathscr{E}$

where $\mathcal{A}^C \subset \mathrm{E}$ is the complement $\{e \in \mathrm{E} \colon e \notin \mathcal{A}\}$ in E of the subset $\mathcal{A} \subset \mathrm{E}$ [16, p. 2], [27, pp. 17-18]. The σ-field $\mathscr{E}$ may also called a σ-algebra over E [27, p. 18], [32, p. 1].

Definition 13 (Pre-Image Set Function)**.** Given a mapping $m \colon \mathrm{E} \longrightarrow \mathrm{L}$ defined between *any* sets E and L, the pre-image of a *subset* $\mathcal{A} \subset \mathrm{L}$ is a *subset* $m^{\natural}(\mathcal{A}) \subset \mathrm{E}$ of E given by:

$$m^{\natural}(\mathcal{A}) \;=\; \{e \in \mathrm{E} \colon m(e) \in \mathcal{A}\} \subset \mathrm{E} \tag{1}$$

where we use the becaudro ($^{\natural}$) to denote pre-image instead of the usual inverse symbol ($^{-1}$) to avoid misinterpretation [32, pp. 3-4]. The pre-image set function $m^{\natural} \colon \mathcal{P}(\mathrm{L}) \longrightarrow \mathcal{P}(\mathrm{E})$ is well-defined, where $\mathcal{P}(L)$ denotes the power set $\{\mathcal{A} \colon \mathcal{A} \subset \mathrm{L}\}$ of L.

When a signal detection system classifies instances of electromagnetic radiation as either containing signals or not, the subset {*instances of electromagnetic radiation* classified *as containing signals amidst noise*} of the event set is the pre-image of a *singleton* subset {*signals amidst noise*} of the label set {*signals amidst noise, noise alone*}.

Definition 14 (Probability Measure)**.** Given a σ-field $\mathscr{E}$ over an *event* set E, a mapping $P \colon \mathscr{E} \to [0,1]$ is a probability measure on $\mathscr{E}$, or, in other words, P is said to be *measurable* with respect to the σ-field $\mathscr{E}$, when the following hold true [29, pp. 41-42]:

(i) $P(\mathcal{E})$ is defined for each event $\mathcal{E} \in \mathscr{E}$,

(ii) $P(\mathrm{E}) = 1$, and

(iii) given any *countable* collection $\{\mathcal{E}_\lambda \in \mathscr{E}\}_{\lambda\in\Lambda}$ of events such that

$\mathcal{E}_\lambda \cap \mathcal{E}_\mu = \varnothing, \quad \forall\ \mu, \lambda \in \Lambda \ \ni\ \mu \neq \lambda$:

$$P\left(\bigcup_{\lambda\in\Lambda} \mathcal{E}_\lambda\right) = \sum_{\lambda\in\Lambda} P(\mathcal{E}_\lambda) \tag{2}$$

Note that a given probability measure $P\colon \mathscr{E} \longrightarrow [0,1]$ may be measurable with respect to other σ-fields besides $\mathscr{E}$, and that pre-images under a probability measure P of all subsets $\mathcal{A} \subset [0,1]$ are *measurable* sets; i.e., they are events in the σ-field over which P is defined.

Definition 15 (Class Prevalence, Prior Probability)**.** Given the following:

(a) a finite index set Λ with cardinality $\mathbf{Card}(\Lambda) = \mathrm{n}$,

(b) a label set L with cardinality $\mathbf{Card}(\mathrm{L}) = \mathrm{n} = \mathbf{Card}(\Lambda)$,

(c) an event set E partitioned by L into classes $\{\mathcal{E}_1, \ldots, \mathcal{E}_\mathrm{n}\}$ satisfying $\bigcup_{\mathrm{j}\in\Lambda} \mathcal{E}_\mathrm{j} = \mathrm{E}$,

(d) a σ-field $\mathscr{E}$ over E such that $\{\mathcal{E}_\mathrm{j}\}_{\mathrm{j}\in\Lambda} \subset \mathscr{E}$, and lastly,

(e) a probability measure $P\colon \mathscr{E} \rightarrow [0,1]$ defined on $\mathscr{E}$,

the class prevalence $\mathrm{p_j}$ for a particular class $\mathcal{E}_\mathrm{j}$ is given by $\mathrm{p_j} = P(\mathcal{E}_\mathrm{j})$. Note that $\mathrm{p_j}$ is also called the *a priori* probability—a.k.a. the *prior probability*—that a given elementary event $e \in \mathrm{E}$ will be contained in class $\mathcal{E}_\mathrm{j}$, for some $\mathrm{j} \in \Lambda$. Since $\{\mathcal{E}_\mathrm{j}\}_{\mathrm{j}=1}^{\mathrm{n}}$ is a partition of E and the probability measure P satisfies $P(\mathrm{E}) = 1 = P\left(\bigcup_{\mathrm{j}=1}^{\mathrm{n}} \mathcal{E}_\mathrm{j}\right)$, then by Definition 14 above, we must have $\sum_{\mathrm{j}=1}^{\mathrm{n}} P(\mathcal{E}_\mathrm{j}) = 1 = \sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{p_j}$.

Theorem 1 (Bayes Theorem)**.** *Given a probability measure* $P\colon \mathscr{E} \longrightarrow [0,1]$ *defined on a* σ*-field* $\mathscr{E}$ *over an event set* E*, and any two events* $\mathcal{X}, \mathcal{Y} \in \mathscr{E}$*, the conditional probabilities* $P(\mathcal{X}|\mathcal{Y})$ *and* $P(\mathcal{Y}|\mathcal{X})$ *have the following scalar relationship:*

$$\begin{aligned} P(\mathcal{X}|\mathcal{Y}) &= \frac{P(\mathcal{X} \cap \mathcal{Y})}{P(\mathcal{Y})} \\ &= \frac{P(\mathcal{Y} \cap \mathcal{X})}{P(\mathcal{Y})} \\ &= \frac{P(\mathcal{Y}|\mathcal{X})P(\mathcal{X})}{P(\mathcal{Y})} \\ &= \left[\frac{P(\mathcal{X})}{P(\mathcal{Y})}\right] P(\mathcal{Y}|\mathcal{X}) \end{aligned} \tag{3}$$

whenever $P(\mathcal{Y}) \neq 0$*; however, if* $P(\mathcal{Y}) = 0$*, then* $P(\mathcal{X}|\mathcal{Y}) = 0, \quad \forall\ \mathcal{X} \in \mathscr{E}$ [33, p. 68].

Definition 16 (Class-Conditional Probability)**.** Given the following:

(a) a classification system $A\colon \mathrm{E} \longrightarrow \mathrm{L}$ with event set domain E and label set range L;

(b) an finite index set Λ satisfying $\mathbf{Card}(\mathrm{L}) = \mathrm{n} = \mathbf{Card}(\Lambda)$,

(c) a σ-field $\mathscr{E}$ over E containing at least the following events:

1. all classes in the partition $\bigcup\limits_{\mathrm{j}\in\Lambda} \mathcal{E}_{\mathrm{j}} = \mathrm{E}$ induced L on E; and
2. all pre-images $A^{\natural}_{\theta}\big(\{\ell_{\mathrm{i}}\}\big) \subset \mathrm{E}$ of *singleton* label subsets $\{\ell_{\mathrm{i}}\} \subset \mathrm{L}$;

(d) a probability measure $P\colon \mathscr{E} \rightarrow [0,1]$ on $\mathscr{E}$; and

(e) a certain class $\mathcal{E}_{\mathrm{j}}$ with *non-zero* prior probability $\mathrm{p}_{\mathrm{j}} = P\big(\mathcal{E}_{\mathrm{j}}\big) \neq 0$ for some $\mathrm{j} \in \Lambda$,

the class-conditional probability $\mathrm{q}_{\mathrm{i}|\mathrm{j}}(A)$ is the *conditional* probability that A assigns a certain label $\ell_{\mathrm{i}} \in \mathrm{L}$ to an elementary event $e \in \mathcal{E}_{\mathrm{j}}$, and is given by:

$$\begin{aligned} \mathrm{q}_{\mathrm{i|j}}(A) &= P\Big(e \in A^{\natural}[\{\ell_{\mathrm{i}}\}] \;\Big|\; e \in \mathcal{E}_{\mathrm{j}}\Big) \\ &= \frac{P\Big(A^{\natural}[\{\ell_{\mathrm{i}}\}] \bigcap \mathcal{E}_{\mathrm{j}}\Big)}{P\big(\mathcal{E}_{\mathrm{j}}\big)} \;\;,\;\; \mathrm{i,j} \,=\, 1,2,3,\ldots,\mathrm{n} \end{aligned} \tag{4}$$

For a class $\mathcal{E}_{\mathrm{j}}$ with prior probability $P\big(\mathcal{E}_{\mathrm{j}}\big) \,=\, 0$, the class-conditional probabilities conditioned on class $\mathcal{E}_{\mathrm{j}}$ are given by $\mathrm{q}_{\mathrm{i|j}}(A) \,=\, 0, \quad \forall \;\; \mathrm{i} \,=\, 1,2,3,\ldots,\mathrm{n}$. A class-conditional probability may take on any value in $[0,1]$, so for each i and j, the *class-conditional* probability $\mathrm{q}_{\mathrm{i|j}}(A)$ is a well-defined probability measure; therefore, by Definition 14 above, we have $\displaystyle\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{q}_{\mathrm{i|j}}(A) \,=\, 1, \quad \forall \;\; \mathrm{j} \,=\, 1,2,3,\ldots,\mathrm{n}$ [29, p. 54].

Assumption 1 (Independence of Class Prevalence and Class-Conditional Probabilities)**.** *Given an* n*-class classification system* A*, any index pair* (i,j), $\mathrm{i},\mathrm{j} \,=\, 1,\ldots,\mathrm{n}$*, and any index* $\mathrm{k} \,=\, 1,\ldots,\mathrm{n}$ *such that class* $\mathcal{E}_{\mathrm{k}}$ *satisfies* $P\big(\mathcal{E}_{\mathrm{k}}\big) \neq 0$*, the set* $\big\{\mathrm{q}_{\mathrm{i|j}}(A), \mathrm{p}_{\mathrm{k}}\big\}$ *of any class-conditional probability* $\mathrm{q}_{\mathrm{i|j}}(A)$ *and any* non-zero *class prevalence* p_{k} *is independent.*

A class-conditional probability might be the likelihood that a signal detection classification system A will label an instance of electromagnetic radiation as class $\mathcal{E}_1$, where this label indicates the presence of signals amidst noise, given that it actually belongs to class $\mathcal{E}_2$ (e.g., the instance observed by the sensor actually contains only noise):

$$\begin{aligned} \mathrm{q}_{1|2}(A) &= P\Big(e \in A^{\natural}[\{\ell_1\}] \;\Big|\; e \in \mathcal{E}_2\Big) \\ &= \left[\frac{P\Big(A^{\natural}[\{\ell_1\}]\Big)}{\mathrm{p}_2}\right] \; P\Big(e \in \mathcal{E}_2 \;\Big|\; e \in A^{\natural}[\{\ell_1\}]\Big) \end{aligned} \tag{5}$$

where the last result is provided by Theorem 1 above.

One result of Assumption 1 would be that as the class prevalence p_2 changes, the probability $P\Big(A^{\natural}[\{\ell_1\}] \cap \mathcal{E}_2\Big)$ must be scaled by exactly the same scalar as is p_2. To visualize this, imagine the event set E as a unit square, with area representing probability. As class prevalences change, so do the sizes of the events within E which they define, so as p_2 changes, the size of the event $\mathcal{E}_2 \subset \mathrm{E}$ changes in exact proportion; Assumption 1 then implies that the size of the event *intersection* $A^\natural\big(\{\ell_1\}\big) \cap \mathcal{E}_2$ must also change such that its area is scaled by the exact same scalar as is the event $\mathcal{E}_2$.

There are several statistical methods available to test the validity of independence between two populations whose distributions are not both known, such as Kendall's Tau [11, pp. 404-405]. The null hypothesis of this particular non-parametric test is *no association* or dependence between the populations [33, p. 816].

To the *user* of a classification system A, the conditional probability $P\Big(e \in \mathcal{E}_\mathrm{j} \;\Big|\; e \in A^\natural\big[\{\ell_\mathrm{i}\}\big]\Big)$ may be of far greater interest than the *class*-conditional probability $P\Big(e \in A^\natural\big[\{\ell_\mathrm{i}\}\big] \;\Big|\; e \in \mathcal{E}_\mathrm{j}\Big)$; however, the set $\Big\{\mathrm{q}_{\mathrm{i|j}}(A)\Big\}_{\mathrm{i,j}=1}^{\mathrm{n}}$ of class-conditional probabilities for A is information by which the system may be judged prior to use, since even if Assumption 1 holds and *class*-conditional probabilities do not change with class prevalences, the class prevalences themselves, such as that in the formula:

$$P\Big(e \in \mathcal{E}_\mathrm{j} \;\Big|\; e \in A^\natural\big[\{\ell_\mathrm{i}\}\big]\Big) = \left[\frac{\mathrm{p_j}}{P\big(A^\natural\big[\{\ell_\mathrm{i}\}\big]\big)}\right] \mathrm{q}_{\mathrm{i|j}}(A)$$

may change from moment to moment, even *while* classification occurs.

Definition 17 (Conditional Probability Matrix)**.** Given a set $\Big\{\mathrm{q}_{\mathrm{i|j}}(A_\theta)\Big\}_{\mathrm{i,j}=1}^{\mathrm{n}}$ of class-conditional probabilities for a classification system A_θ, the conditional probability matrix is given by $\Big[\mathbf{Q}_{A_\theta}\Big]_{\mathrm{ij}} = \mathrm{q}_{\mathrm{i|j}}(A_\theta)$.

A collection of conditional probability matrices for various classification systems may be represented by $(\mathrm{n}^2-\mathrm{n})$-dimensional vectors in the Cartesian product $[0,1]^{\mathrm{n}^2-\mathrm{n}} \subset \mathbb{R}^{\mathrm{n}^2-\mathrm{n}}$. If one considers a family $\mathbb{A}_\Theta = \{A_\theta \colon \theta \in \Theta\}$ of classification systems over a threshold set Θ of interest with only *continuous* parameters, a continuous $(\mathrm{n}^2-\mathrm{n})$-dimensional surface may then be constructed by infinitesimal variations of these parameters; in practice, however, such continuous curves may only be estimated by a finite number of ROC vector estimates representing classification systems in the family $\mathbb{A}_\Theta$.

The most common method of representing an estimate of a class-conditional probabilities is by calculating a transpose stochastic confusion matrix from experimental results. There are, of course, other methods of obtaining class-conditional probability estimates, and the distribution of ROC vectors may even be defined statistically; Assumption 1 then allows these distributions to be treated separately from any distributions attributed to class prevalences.

To illustrate the calculation of a transpose stochastic confusion matrix, consider a 2×2 *contingency matrix* of raw results for a binary classification experiment (or observational study) with a finite number of classification results and *a priori* (or *a posteriori*, in the case of an observational study) knowledge of class populations for all exemplars classified. Such a matrix displays a simple count of the numbers of each type of decision, including both correct and incorrect decisions, with correct decision counts along the diagonal and with columns corresponding to the the truth, as shown in Table 1.

Table 1: Two-Class Contingency Matrix.

$$\begin{array}{rc} \text{Contingency Matrix} & \begin{array}{cc} \text{Actual Class: 1} & \text{Actual Class: 2} \end{array} \\ \begin{array}{r} \text{Labeled Class: 1} \\ \text{Labeled Class: 2} \end{array} & \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \end{array}$$

Here, class 1 is the so-called *positive* or target class, and class 2 the *negative*; hence, TP or the *true positive* count is how many exemplars from class 1 were correctly labeled, and FN or the *false negative* count is how many were not, etc. [10, pp. 69-71].

Estimates of class-conditional probabilities may be formed by dividing each element of a class-specific column in the contingency matrix by the total number of classified exemplars from that class. With M_1 and M_2 exemplars from Classes 1 and 2, respectively, undergoing classification, we may estimate the class-conditional probabilities from Table 1, as shown in Table 2.

Table 2: Two-Class Confusion Matrix.

$$\begin{array}{lcc} \text{Confusion Matrix} & \text{Actual Class: 1} & \text{Actual Class: 2} \\ \text{Labeled Class: 1} & \frac{TP}{M_1} & \frac{FP}{M_2} \\ \text{Labeled Class: 2} & \frac{FN}{M_1} & \frac{TN}{M_2} \end{array}$$

The result is a *transpose stochastic confusion matrix*, such that the sum of each column is one. It is worth mentioning that some authors prefer the proper stochastic presentation, but for the purposes of this thesis, the transpose stochastic is more convenient [6, pp. 8-9]. Also, the term “confusion matrix” sometimes means the contingency matrix denoted above, and a normalized form of the contingency matrix as illustrated above (that which *we* term a confusion matrix) may be specified as a “confusion rate matrix” or “confusion ratio matrix” to avoid confusion with the non-normalized form [8, p. 3], [9, p. 2]. Due to its transpose stochastic nature, the information contained in a 2×2 confusion matrix of this type may be presented as a coordinate pair comprised of one entry from each column, which may then be plotted on a unit square [30, pp. 26-28].

Assumption 2 (Acceptable Class-Conditional Probability Estimates)**.** *Without regard to the method of obtaining an* estimate $\widehat{\mathbf{Q}_A}$ *of the conditional probability matrix* $\mathbf{Q}_A$ *for a given classification system A, assume that adequate estimation procedures have occurred, such that for all practical purposes, considering* $\widehat{\mathbf{Q}_A}$ *approximately equal to the matrix* $\mathbf{E}[\,\mathbf{Q}_A\,]$ *of expected values of the elements of* $\widehat{\mathbf{Q}_A}$ *results in no appreciable error; i.e., we may substitute* $\mathbf{E}[\,\mathbf{Q}_A\,] \approx \widehat{\mathbf{Q}_A}$ *whenever it is convenient to do so.*

Definition 18 (ROC Manifold, ROC Curve)**.** Given an n-class classification problem, the convex hull of a continuous collection of ROC vectors estimates plotted in $(\mathrm{n}^2 - \mathrm{n})$-dimensional space is often termed a ROC curve (for a two-class scenario) or a ROC manifold [30]. The ROC Convex Hull is abbreviated ROCCH.

If constructing the ROCCH was simple, then comparing only classification systems whose points lie on hull might save time, since no points within the hull interior could possibly represent classification systems superior to those on the hull under any circumstances [24], [30]. Such considerations would reduce the number of classification systems to compare and contrast; however, since the simplicity of ROCCH calculation, and thus the amount of time to be possibly saved, is questionable, the *method* of comparison would seem to be *far* more important than saving time during such a comparison, depending, of course, on the possible applications of the classification system. Except for time-saving purposes, such geometrical concepts have limited utility under decision-theoretical constructs, yet the ROCCH, especially in its binary form as the ROC curve, has played a huge role in ROC analysis for many years, and are therefor worthy of mention; however, they are not actually *necessary* considerations within the framework of risk calculation; therefore, this thesis will refer to them only as auxiliary concepts.

Definition 19 (Cost Matrix)**.** A cost matrix given by $[\mathbf{C}]_{\mathrm{ij}} = \mathrm{c}_{\mathrm{i|j}}$ is an $\mathrm{n} \times \mathrm{n}$ matrix of real numbers representing costs or *losses* $\mathrm{c}_{\mathrm{i|j}}$ specific to events $\left(e \in A^{\natural}\big[\{\ell_{\mathrm{i}}\}\big] \;\middle|\; e \in \mathcal{E}_{\mathrm{j}}\right)$, i.e., classification system A assigns label ℓ_{i} to an elementary event e when it is *actually* an element of class $\mathcal{E}_{\mathrm{j}}$, whose *class-conditional* probability holds the exact same (i, j)-position in the conditional probability matrix $\left[\mathbf{Q}_A\right]_{\mathrm{ij}} = \mathrm{q}_{\mathrm{i|j}}(A)$. These costs *may* be positive or negative, but most often, the sum of off-diagonal entries in any column is *greater* than the diagonal entry itself, indicating that it is *better* (i.e., less costly) to classify something as what it *actually is* rather than anything else [5, pp. 24-25]. This matrix may also be called a "payoff" matrix, so its meaning is almost completely subjective [6, p. 16]. One common form is the so-called "zero-one" *transpose stochastic* cost matrix, with all zeroes on the diagonal and each column summing to one; however, it is not necessary to restrict the cost matrix to such a form [5, p. 26], [32, p. 7].

Assumption 3 (Fixed Costs)**.** *All elements of a given cost matrix* $\boldsymbol{C}$ *are fixed.*

This is a necessary assumption, because costs often are the result of human reasoning, which is very unpredictable; therefore, it is easier to simply choose different possible cost regimes and perform risk calculations under each scenario.

Assumption 4 (Independence of Class-Conditional Costs and Probabilities)**.** *Given an* n*-class classification system* A *and any index pairs* (i, j)*,* $\mathrm{i}, \mathrm{j} = 1, \ldots, \mathrm{n}$ *and* (h, k)*,* $\mathrm{h}, \mathrm{k} = 1, \ldots, \mathrm{n}$*, the set* $\left\{\mathrm{q}_{\mathrm{i|j}}(A), \mathrm{c}_{\mathrm{h|k}}\right\}$ *consisting of any class-conditional probability* $\mathrm{q}_{\mathrm{i|j}}(A)$*, and any cost* $\mathrm{c}_{\mathrm{h|k}}$*, is independent.*

Since costs are subjectively defined, it may be possible to envision a scenario where the likelihood of making a particular type of classification decision has a direct impact on the cost of such decision; however, it should therefore also be possible to define scenarios where costs do not change as estimates of ROC information change.

Definition 20 (Prevalence Matrix)**.** Given an set $\left\{\mathrm{p_j}\right\}_{\mathrm{j=1}}^{\mathrm{n}}$ of class prevalences for a classification system A, the prevalence matrix $\mathbf{P}$ is an $\mathrm{n \times n}$ *stochastic* matrix with each row the *same* ordered n-tuple $\mathbf{p}^T$ consisting of the class prevalences $\left\{\mathrm{p_j}\right\}_{\mathrm{j=1}}^{\mathrm{n}}$:

$$\mathbf{p} \equiv \begin{bmatrix} \mathrm{p_1} \\ \vdots \\ \mathrm{p_n} \end{bmatrix}_{\mathrm{n\times 1}} \implies \mathbf{P} = \begin{bmatrix} \mathbf{p}^T \\ \vdots \\ \mathbf{p}^T \end{bmatrix}_{\mathrm{n\times n}} = \begin{bmatrix} \mathrm{p_1} & \cdots & \mathrm{p_n} \\ & \vdots & \\ \mathrm{p_1} & \cdots & \mathrm{p_n} \end{bmatrix}_{\mathrm{n\times n}} \tag{6}$$

Assumption 5 (Independence of Class Prevalence and Class-Conditional Costs)**.** *Given an* n*-class classification system* A*, any index pair* $\mathrm{(i,j),\ i,j = 1, \ldots, n}$*, and any index* $\mathrm{k = 1, \ldots, n}$*, the set* $\left\{\mathrm{c_{i|j}, p_k}\right\}$ *consisting of any cost* $\mathrm{c_{i|j}}$ *and any prior class probability* $\mathrm{p_k}$*, is independent.*

Since the definition of cost is purely subjective, it is certainly possible that the individual costs of making classification decisions may be independent of the class prevalences. For example, in signal detection, this might be like assuming that it is always equally costly to assume an instance of electromagnetic radiation contains noise alone, given that it actually contains a signal, since the binary nature of the setup seems to imply that there is potentially valuable information contained in *any* type of signal.

Definition 21 (Matrix Hadamard Product)**.** Given any two matrices $\mathbf{U}$ and $\mathbf{V}$ of the same size, the binary Hadamard Matrix Operator $\odot$ forms a new matrix $\mathbf{U} \odot \mathbf{V}$ of the same size. A typical element of the resultant matrix is given by:

$$[\mathbf{U} \odot \mathbf{V}]_{\mathrm{ij}} = \mathrm{u_{ij}\, v_{ij}} = \mathrm{v_{ij}\, u_{ij}} = [\mathbf{V} \odot \mathbf{U}]_{\mathrm{ij}} \tag{7}$$

Definition 22 (Frobenius Dot Product)**.** Given any two matrices $\mathbf{U}$ and $\mathbf{V}$ of size $\mathrm{s} \times \mathrm{r}$, the Matrix Dot Product Operator $\langle\ ,\ \rangle_F$ performs the following reflexive binary operation:

$$\langle \mathbf{U}, \mathbf{V} \rangle_F \;=\; \sum_{\mathrm{i}=1}^{\mathrm{s}} \left(\sum_{\mathrm{j}=1}^{\mathrm{r}} \mathrm{v}_{\mathrm{ij}}\, \mathrm{u}_{\mathrm{ij}} \right) \;= \sum_{\mathrm{j}=1}^{\mathrm{r}} \left(\sum_{\mathrm{i}=1}^{\mathrm{s}} \mathrm{u}_{\mathrm{ij}}\, \mathrm{v}_{\mathrm{ij}} \right) \;=\; \langle \mathbf{V}, \mathbf{U} \rangle_F \tag{8}$$

which is simply the sum of all elements of the Hadamard Matrix Product $\mathbf{U} \odot \mathbf{V}$.

Definition 23 (Standard m-Simplex)**.** Given *any* positive integer m, along with *any* ordered m-tuple $\mathbf{p} \in [0,1]^{\mathrm{m}}$ of non-negative real variables $\mathrm{p_j}$, $\mathrm{j} = 1, 2, 3, \ldots, \mathrm{m}$, the standard m-simplex Δ_{m} is the set [3, p. 568], [20, pp. 149-150]:

$$\Delta_{\mathrm{m}} \;=\; \left\{ \mathbf{p} \in [0,1]^{\mathrm{m}} : \; \sum_{\mathrm{j}=1}^{\mathrm{m}} \mathrm{p_j} \;\leq\; 1 \right\} \tag{9}$$

Figure 1 shows an example of how , for a three-class scenario, a two-dimensional prevalence vector $\mathbf{p} \;=\; \big(\mathrm{p}_1, \mathrm{p}_2\big)$ whose coordinates sum to a number less than 1 may be drawn from Δ_2 . Note that the unspecified value of p_3 is found from the conjunctive equation $\sum_{\mathrm{j}=1}^{3} \mathrm{p_j} \;=\; 1$, as illustrated in Figure 2, where this point lies on the tilted surface of a standard 3-simplex.

Definition 24 ($\mathrm{s} \times \mathrm{r}$ Random Matrix)**.** Given an $\mathrm{s} \times \mathrm{r}$ matrix $\mathbf{B} \;=\; \big[\mathrm{B_{ij}}\big]$ of event sets and an $\mathrm{s} \times \mathrm{r}$ matrix $\mathbf{X}$ of functions $\mathrm{x_{ij}}$ defined on $\mathrm{B_{ij}}$, respectively, $\mathbf{X}$ is an $\mathrm{s} \times \mathrm{r}$ matrix of random variables, or a *random matrix*, when the codomain of each function $\mathrm{x_{ij}} : \mathrm{B}_{\mathrm{ij}} \longrightarrow \mathbb{R}$ is the set $\mathbb{R}$ of real numbers [33, p. 73].

There is no stipulation as to what type of event set a random variable may be defined upon; therefore, any or all of the event sets in a matrix $\mathbf{B}$ of event sets may be either

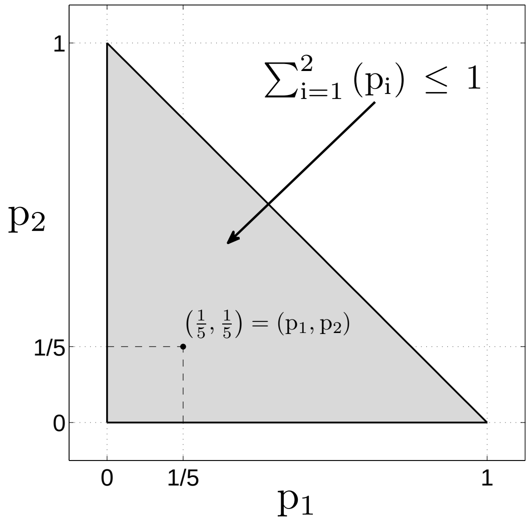

Figure 1: $\Delta_2 \equiv \left\{ \mathbf{p} \in [0,1]^2 : \sum_{j=1}^{2} p_j \leq 1 \right\}$.

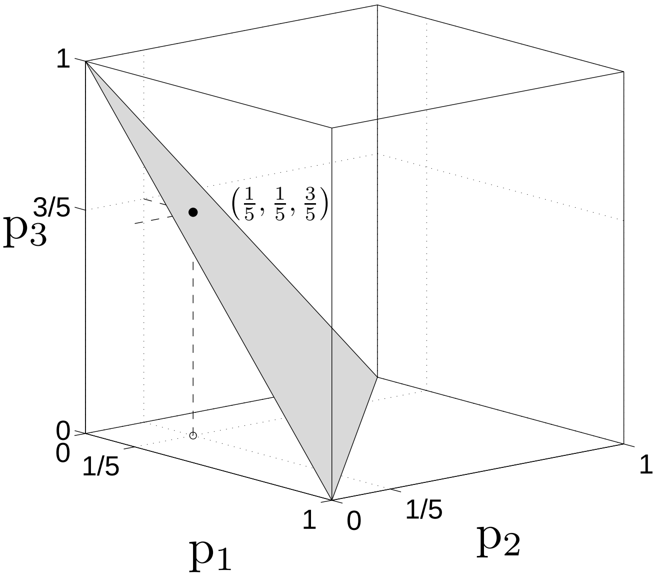

Figure 2: $\Delta_3 \equiv \left\{ \mathbf{p} \in [0,1]^3 : \sum_{j=1}^{3} p_j \leq 1 \right\}$.

discrete or continuous. Given an n-class classification system A and the corresponding real-valued matrices $\mathbf{Q}_A$, $\mathbf{C}$, and $\mathbf{P}$ introduced in Definitions 17, 19, and 20 above, respectively, note that each is a random matrix defined on a matrix of event sets.

With regard to the matrices $\mathbf{Q}_A$ and $\mathbf{P}$, it is also evident that there are exactly n random variables that are functions of only $(\mathrm{n}-1)$ of the random variables inhabiting the same column or row, due to the respective transpose stochastic and stochastic natures of these matrices. For example, since the prevalence matrix $\mathbf{P}$ is simply the same random vector $\mathbf{p}$ arrayed next to itself n times, the *definitions* of all n of these functionally dependent variables are exactly the same; similarly, of the remaining $(\mathrm{n}^2-\mathrm{n})$ random variables that *could* be non-constant, there are actually only $(\mathrm{n}-1)$ *unique* random variables. It will become apparent later why this notation is used; it is sufficient for now to notice that any joint distribution defined for $\mathbf{P}$ will be a function of the same $(\mathrm{n}-1)$ unique random variables that populate each of its rows, as will any joint distribution defined for a given column of $\mathbf{Q}_A$. The respective stochastic and transpose stochastic natures of these matrices means that an entire row or column vector of random variables will be jointly distributed over a standard $(\mathrm{n}-1)$-simplex (see Definition 23 above), since, for example, the n^{th} entry $\mathrm{p_j}$ randomly drawn in each row vector $\mathbf{p}$ in the prevalence matrix $\mathbf{P}$ is a function $\mathrm{p_n} = 1-\sum_{\mathrm{j}=1}^{\mathrm{n}-1}\mathrm{p_j}$ of the other $(\mathrm{n}-1)$ random variables in the row which are randomly drawn before it.

1.3 Problem Statement

Before proceeding further, let us motivate the need for the preceding definitions by means of the following situation. Imagine a stockbroker analyzing the contents of a certain client's portfolio, implementing an algorithm that classifies stocks as either *buy*, *sell*, or *hold.* Although unknown to the broker or the directors of the corporations whose stocks she analyzes, there are a set of seemingly insignificant factors that, when occurring simultaneously, create severe danger of financial ruin for many of these corporations. Her classification system was created to detect just such problems, however, and it reports that 85% of stocks in this particular portfolio are *sell* stocks—i.e., stocks that ought to be sold immediately. Since the broker has never seen numbers for the *sell* class greater than 10%, she begins to question the results, and therefore does not immediately sell those stocks. Time ticks by, and it becomes more readily apparent to the corporations and the broker that the stocks are highly over-valued, and the window of opportunity to sell with minimal loss shrinks away overnight.

If the broker in this case had been informed beforehand that the classification system she used had been selected and tuned specifically to the cost structure dictated by her management, and that the distribution of stock class prevalences provided for the possibility of unknown factors causing a change in stock class prevalences, she might have had more confidence in the classification system, and then acted immediately to avoid losing more money for her client, because her risk was already minimized by acting on the results of the classification system.

Although the scenario above might be unrealistic, there are many classification situations which entail potentially much greater costs, e.g., from the loss of life (military applications are just one). However, many popular methods of comparing classification systems to one another do not consider the whole picture of risk—i.e., the costs *and* class prevalences *in addition* to an estimate of the class-conditional probabilities. In addition to these oversights, and due to the fact that volume under a ROC surface (VUS) in a three-class case would have six dimensions, visualization of geometric surfaces becomes impossible, so ignoring more than just one of the entries per column of a conditional probability matrix estimate is also sometimes chosen as an alternative [21]. Most attempts to generalize geometric concepts to the general n-class case choose to ignore either the class prevalences or the costs [7], [9], [35]. If Assumptions 1, 2, 3, 4, and 5 are relatively safe assumptions to make, then the concept of risk offers the opportunity a much more robust form of ROC analysis; i.e., one which considers many more of the *characteristics* of the *operating* environment in which the *receiver* of information resides.

II. Review of Related ROC Analysis Topics

The monograph by Egan serves as a starting point for modern binary ROC analysis. It contains much of the terminology and geometry still in use today, as well as a framework for risk calculations [6, p. 16]. Based on his work, for a given classification system A and accompanying conditional probability, cost, and prevalence matrices $\mathbf{Q}_A$, $\mathbf{C}$, and $\mathbf{P}$ as given in Definitions 9, 17, 19, and 20 of Chapter I, respectively, we define the risk $R(A)$ of a classification system A (suppressing notational dependence on A) as :

$$R \;=\; \Big\langle \mathbf{Q}\,, \big(\mathbf{C} \odot \mathbf{P}\big) \Big\rangle_F \tag{10}$$

with Matrix Hadamard Product $\odot$ and Frobenius Dot Product $\langle\, ,\, \rangle_F$ as defined in Chapter I, Definitions 21 and 22, respectively. Egan notes that (10) gives the expected cost over a sufficiently large number of trials; therefore, from this point onward, we shall assume, as in Chapter I, Assumption 2, that such is the case [6, pp. 16-17].

2.1 Two-Class ROC Analysis

Assume the following notation of a transpose stochastic confusion matrix for some binary classification system A (again, suppressing notational dependence on A):

$$\begin{bmatrix} \widehat{q_{1|1}} & \widehat{q_{1|2}} \\ \widehat{q_{2|1}} & \widehat{q_{2|2}} \end{bmatrix} \;=\; \begin{bmatrix} tpr & fpr \\ fnr & tnr \end{bmatrix} \tag{11}$$

where class 1 assumes the role of the so-called *positive* or target class, and class 2 is the *negative* or non-target class, thereby leading to the abbreviations for true positive rate (*tpr*), false negative rate (*fnr*), false positive rate (*fpr*), and true negative rate (*tnr*) in common use today [10], [32]. Since this matrix is transpose stochastic, its information may be represented by only one entry from each column. Although there are four possible ways to do this, the common way is to plot (fpr, tpr) as in Figure 3, so that the coordinate representation of a perfect classifier is at $(0, 1)$. In this coordinate system, maximal area beneath the lines connecting a plotted point for a given classification system to the corners $(0, 0)$ and $(1, 1)$ is seen as desirable [8, pp. 108-110].

Based on this geometrical frame of reference, one of the most popular means of evaluating classification system effectiveness is by the Area Under the ROC Curve (AUC) performance measure, which calculates geometrically the area under the convex hull of a collection of ROC vector estimates plotted in this way to represent a family of binary classification systems [10], [13], [26]. Instead of analyzing *collections* of ROC vectors, consider the case with just one plotted ROC vector [8, pp. 108-110], as in Figure 3.

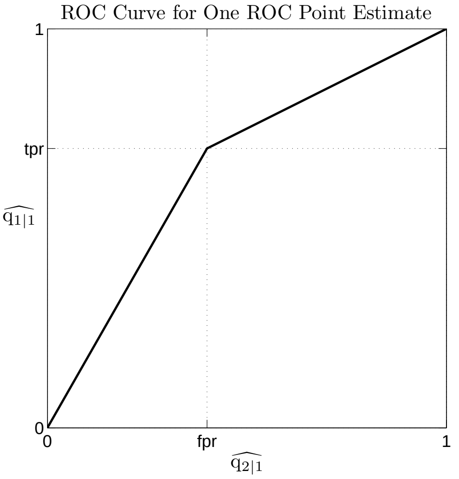

Figure 3: ROC Curve for One ROC Point Estimate.

The area under this ROC curve is simply the sum of a square and two triangles:

$$\begin{aligned}
\text{AUC} &= (1 - fpr)(tpr) \;+\; \frac{(fpr)(tpr)}{2} + \frac{(1 - fpr)(1 - tpr)}{2} \\
&= \frac{\big[2(1 - fpr)(tpr)\big] \;+\; \big[(fpr)(tpr)\big] \;+\; \big[(1 - fpr)(1 - tpr)\big]}{2} \\
&= \frac{2\big[tpr \;-\; (fpr)(tpr)\big] \;+\; (fpr)(tpr) \;+\; 1 \;-\; tpr \;-\; fpr \;+\; (tpr)(fpr)}{2} \\
&= \frac{2(tpr) \;-\; 2(fpr)(tpr) \;+\; (fpr)(tpr) \;+\; 1 \;-\; tpr \;-\; fpr \;+\; (tpr)(fpr)}{2} \\
&= \frac{tpr \;+\; (1 \;-\; fpr)}{2} \\
&= \frac{tpr \;+\; tnr}{2}
\end{aligned} \tag{12}$$

where the last observation is made possible by the conjunctive equation $fpr \;+\; tnr \;=\; 1$ pertaining to the left columns in (11) above. Now, if we assume equal class prevalences $M_1 \;=\; M \;=\; M_2$ we may write:

$$\begin{aligned}
\text{AUC} &= \frac{tpr \;+\; tnr}{2} \\
&\equiv \frac{\frac{TP}{M_1} \;+\; \frac{TN}{M_2}}{2} \\
&= \frac{\frac{TP}{M} \;+\; \frac{TN}{M}}{2} \\
&= \frac{TP \;+\; TN}{2M} \\
&= \frac{TP \;+\; TN}{M_1 \;+\; M_2} \\
&\equiv \text{Accuracy}
\end{aligned} \tag{13}$$

Accuracy is related to risk through the AUC, as seen when we calculate the approximate risk $R \;\approx\; \Big\langle \widehat{\mathbf{Q}} \,,\, \big(\mathbf{C} \odot \mathbf{P}\big) \Big\rangle_F$ (per Assumption 2, Chapter I) indicated by (10) for a zero-one cost matrix $\mathbf{C} \;=\; \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ under equal class prevalences:

$$
\begin{aligned}
R &\approx \left\langle \widehat{\mathbf{Q}}\,,\,(\,\mathbf{C}\odot\mathbf{P})\right\rangle_F \\
&= \left\langle \begin{bmatrix} tpr & fpr \\ fnr & tnr \end{bmatrix}, \left(\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \odot \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \right) \right\rangle_F \\
&= \left\langle \begin{bmatrix} tpr & fpr \\ fnr & tnr \end{bmatrix}, \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix} \right\rangle_F \\
&= \frac{(0\ +\ fpr)\ +\ (fnr\ +\ 0)}{2} \\
&= \frac{(1\ -\ tnr)\ +\ (1\ -\ tpr)}{2} \\
&= 1\ -\ \frac{tpr\ +\ tnr}{2} \\
&= 1\ -\ \text{AUC}
\end{aligned}
\tag{14}
$$

where the last relation again is made possible by the conjunctive equations from (11) above; therefore, by (12) above, the risk R for a zero-one cost matrix and equal class prevalences is simply (1 $-$ Accuracy) under the same assumptions, and (1 $-$ AUC) in general for a ROC curve with only one point. It is interesting to note that if the coordinate pair used to represent the information of the transpose stochastic matrices in (11) were (fpr, fnr) instead of (fpr, tpr), the calculation in (14) would yield $R \approx$ AUC.

Neither the AUC nor Accuracy consider costs, but unlike Accuracy, the AUC also does not consider class prevalences in its calculation, and so the AUCs for two very different classification systems may be equal, as shown in Figure 4.

There is some merit to the idea that the conventional formula for Accuracy considers class prevalences, but it still ignores the costs, and for that reason is incomplete as a measure of risk [10], [14], [19], [25]. It is also not robust to changes in class prevalences when extended to a classification system with more than 2 classes. There are quite a few other performance measures related to Accuracy or the AUC which we shall not mention,

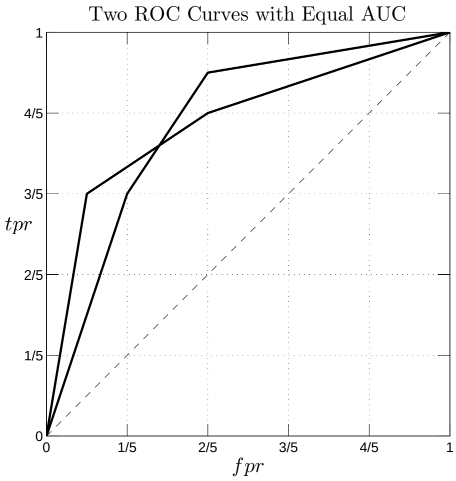

Figure 4: Two ROC Curves with Equal AUC.

due to the similarity of their weaknesses to changes in class prevalences and differing costs. Proceeding in this manner, it becomes apparent that any other ROC analysis calculation based on Accuracy or the AUC is equivalent to a risk calculation with certain restrictions on the values of the cost and prior information.

In general, it appears that none of the binary ROC analysis methods in popular use today truly utilize the significant Bayesian inputs of costs and prior probabilities.

2.2 Multi-Class ROC Analysis

Extending classical ROC analysis beyond the realm of binary classification is difficult. Some authors have proposed using only one entry per column of a 3×3 transpose stochastic confusion matrix, eliminating most of the ROC information, and explicitly considering neither costs nor class prevalences in their calculations [21, pp. 80-82], [22, p. 3441] [34, p. 4]. More recent approaches consider *either* the costs *or* the priors as one of the parameters in the threshold set, ignoring the effect of the other, or suggest plotting different curves for each *pair* of classes as done in binary ROC analysis [9], [12], [36].

Due to the conjunctive equations accompanying any conditional probability matrix, the Volume Under the Surface (VUS) for an n-class scenario is only a true extension of the AUC when it is $(n^2 - n)$-dimensional, however, some authors, in order to produce visible surfaces, plot only a 3-dimensional surface for a 3-class system [4], [21], [22]. Some, who realize the weakness such a scheme entails, allude to the calculation of risk; however, the calculation is not performed, because, for example, the need to assume unknown costs is deemed important. In general, breaching Assumptions 1, 2, 3, 4, *and* 5 from Chapter I is never mentioned as a cause for not calculating the risk [8], [9].

2.3 Need for a ROC Risk Functional

The world of classical ROC analysis seems to be stuck unnecessarily in a frame of reference that considers geometrical analyses as the gold standard of ROC analysis methods, when in fact, if Assumptions 1, 2, 3, 4, and 5 from Chapter I can be met, the comparison of classification systems by risk comparison is not simpler and more comprehensive. In addition, risk-based comparison of classification systems falls closer to the reach of the ordinary decision-maker, who is usually not involved in obtaining estimates of class-conditional probabilities, but usually *is* responsible for defining costs and may even have some knowledge of class prevalence distributions. When risk-based comparisons of classification systems are implemented, the result is much simpler, as well as more considerate of the crucial role of *both* costs *and* class prevalences in ROC analysis [32].

III. Development and Definition of the ROC Risk Functional

The ROC Functional $f_{\mathbb{A}}$ suggested in [31] and [32] for a family $\mathbb{A}$ of classification systems is a ROC analysis method which minimizes risk. Additionally, it was proposed in [32] to allow the Hadamard Product $\hat{\boldsymbol{\gamma}} = \big(\mathbf{c} \odot \mathbf{p}\big)$ of vectors of costs and class prevalences, constructed in a particular way, to vary over a range $\Gamma = \{\boldsymbol{\gamma} \colon \boldsymbol{\gamma} = \mathbf{c} \odot \mathbf{p}\}$, along with restrictions on cost. This Robust Functional implicitly incorporated Assumptions 1, 2, and 4 from Chapter I, but did not incorporate Assumptions 3 and 5 from that Chapter, which made the problem more difficult. The equation was written as:

$$R\big(f_{\mathbb{A}}, \Gamma\big) = \min_{\mathbf{q} \in Q} \int_{\Gamma} \langle \mathbf{q}, \widehat{\boldsymbol{\gamma}} \rangle \, \mathrm{W}(\boldsymbol{\gamma}) \, \mathbf{d}\boldsymbol{\gamma} \tag{15}$$

where Q is a collection $Q = \{\mathbf{q} : \mathbf{q} \in Q\}$ of ROC vectors corresponding to the family $\mathbb{A}$ of classification systems and $\mathrm{W}(\boldsymbol{\gamma})$ is a *joint* weighting function, of the cost-prior Hadamard Product vector $\widehat{\boldsymbol{\gamma}} = \mathbf{c} \odot \widehat{\mathbf{p}}$, cast either as a probability density function or a belief function [32, p. 6].

In addition to the implicit incorporation of Assumptions 1, 2, and 4 from Chapter I, if we also incorporate Assumptions 3 and 5 from the same Chapter, we may fashion the distributions of costs and priors independently from one another by making over $\mathbf{q}$, $\mathbf{c}$, and $\mathbf{p}$ in (15) above to be the *random matrices* $\mathbf{Q}_A$, $\mathbf{C}$, and $\mathbf{P}$ (see Definitions 24, 17, 19, and 20 from Chapter I). Without explicitly denoting dependence on the classification system A, we define marginal weighting functions $W_{\mathbf{Q}}(\mathbf{Q})$, $W_{\mathbf{C}}(\mathbf{C})$, and $W_{\mathbf{P}}(\mathbf{P})$. Since these marginal distributions are defined for sets of random variables assumed independent from one another, they satisfy the separability condition [33, p. 245]:

$$W_{\mathbf{Q,C,P}}(\mathbf{Q},\mathbf{C},\mathbf{P}) \;=\; W_{\mathbf{Q}}(\mathbf{Q})\, W_{\mathbf{C}}(\mathbf{C})\, W_{\mathbf{P}}(\mathbf{P})$$

We shall now examine possible joint probability density functions $W_{\mathrm{P}}(\mathbf{P})$ on the priors such that the constraints of the conjunctive equation $\sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{p_j} \;=\; 1$ are met.

Note that (15) simply calculates Bayes risk, or the expected value of Equation (10), Chapter II. Without explicitly denoting dependence on the classification system A or functional dependence on the variables in the matrices $\mathbf{Q}_A$, $\mathbf{C}$, and $\mathbf{P}$, we may write:

$$\begin{aligned}
E(R) \;&\equiv\; E\left[\Big\langle \mathbf{Q}\,,\,(\mathbf{C}\odot\mathbf{P})\Big\rangle_F\right] \\
&= \int\!\!\int\!\!\int \langle \mathbf{Q}, (\mathbf{C}\odot\mathbf{P})\rangle_F \, W_{\mathbf{Q,C,P}}\ \mathbf{dQ}\,\mathbf{dC}\,\mathbf{dP} \\
&= \int\!\!\int\!\!\int \sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\sum_{\mathrm{j}=1}^{\mathrm{n}}\Big[\mathrm{q_{i|j}}\big(\mathrm{c_{i|j}}\,\mathrm{p_j}\big)\Big]\right) W_{\mathbf{Q,C,P}}\ \mathbf{dQ}\,\mathbf{dC}\,\mathbf{dP} \\
&= \sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\sum_{\mathrm{j}=1}^{\mathrm{n}}\left[\int\!\!\int\!\!\int \mathrm{q_{i|j}}\,\mathrm{c_{i|j}}\,\mathrm{p_j}\,\big(W_{\mathbf{Q}}W_{\mathbf{C}}W_{\mathbf{P}}\ \mathbf{dQ}\big)\,\mathbf{dC}\,\mathbf{dP}\right]\right) \\
&= \sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\sum_{\mathrm{j}=1}^{\mathrm{n}}\left[\int \mathrm{p_j}\,W_{\mathbf{P}}\int \mathrm{c_{i|j}}W_{\mathbf{C}}\int \mathrm{q_{i|j}}\,W_{\mathbf{Q}}\ \mathbf{dQ}\,\mathbf{dC}\,\mathbf{dP}\right]\right) \\
&= \sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\sum_{\mathrm{j}=1}^{\mathrm{n}}\left[\Big(\int \mathrm{p_j}\,W_{\mathbf{P}}\,\mathbf{dP}\Big)\Big(\int \mathrm{c_{i|j}}W_{\mathbf{C}}\,\mathbf{dC}\Big)\Big(\int \mathrm{q_{i|j}}\,W_{\mathbf{Q}}\,\mathbf{dQ}\Big)\right]\right) \qquad (16)\\
&= \sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\sum_{\mathrm{j}=1}^{\mathrm{n}}\left[\Big(E\big[\mathrm{p_j}\big]\Big)\Big(E\big[\mathrm{c_{i|j}}\big]\Big)\Big(E\big[\mathrm{q_{i|j}}\big]\Big)\right]\right) \\
&= \sum_{\mathrm{i}=1}^{\mathrm{n}}\left[\sum_{\mathrm{j}=1}^{\mathrm{n}}\left(E\big[\mathrm{q_{i|j}}\big]\,\Big[E\big[\mathrm{c_{i|j}}\big]\,E\big[\mathrm{p_j}\big]\Big]\right)\right] \\
&= \Big\langle \mathbf{E}\,[\,\mathbf{Q}\,]\;,\;\Big(\mathbf{E}\,[\,\mathbf{C}\,]\odot\mathbf{E}\,[\,\mathbf{P}\,]\Big)\Big\rangle_F \\
&\approx \Big\langle\;\widehat{\mathbf{Q}}\;,\;\mathbf{C}\odot\mathbf{E}\,[\mathbf{P}]\Big\rangle_F
\end{aligned}$$

where the boldface expected value $\mathbf{E}\big[\,\cdot\,\big]$ denotes a *matrix* $\big[\,E(\cdot)\,\big]_{\mathrm{ij}}$ of expected values, and where we have introduced the notation of integration with respect to a matrix, such that $\int [\,\cdot\,]\ \mathbf{dX}$ denotes the multiple integration operator:

$$\int \ldots\ \ldots \int \ldots \int\int [\,\cdot\,]\ d\mathrm{x}_{11}\, d\mathrm{x}_{12}\ \ldots\ d\mathrm{x}_{1\mathrm{r}}\ \ldots\ \ldots\ d\mathrm{x}_{\mathrm{sr}}$$

with respect to all of the variables in the matrix $\mathbf{X}$ of size $\mathrm{s} \times \mathrm{r}$, such that $\mathbf{dX}$ denotes the product of all differential elements $d\mathrm{x}_{\mathrm{ij}}$ of variables in $\mathbf{X}$, $\forall\ \mathrm{i} \in \{1, 2, 3, \ldots, \mathrm{s}\}$ and $\mathrm{j} \in \{1, 2, 3, \ldots, \mathrm{r}\}$. Note that without Assumptions 1, 2, 3, 4, and 5 from Chapter I, we could not perform this simple dot product calculation for Bayes risk [33, p. 233-246].

3.1 Definition of the ROC Risk Functional

Given a family $\mathbb{A}_\Theta = \{A_\theta \colon \theta \in \Theta\}$ of n-class classification systems of form $A_\theta \colon \mathrm{E} \longrightarrow \mathrm{L}$ over a threshold set Θ, with common cost and prevalence matrices $\mathbf{C}$ and $\mathbf{P}$ and a collection $\left\{\mathbf{Q}_{A_\theta} \colon \theta \in \Theta\right\}$ of conditional probability matrices, as defined in Chapter I, Definitions 11, 10, 19, 20, and 17, respectively, define the ROC Risk Functional (RRF) as a threshold parameter $\theta \in \Theta$ such that the classification system A_θ minimizes Bayes risk over the family $\mathbb{A}_\Theta$ of classification systems:

$$\begin{aligned} \arg \min_{A_\theta \in \mathbb{A}_\Theta} \left\{ E\left[\, R_{A_\theta} \,\right] \right\} \ &\equiv\ \arg \min_{A_\theta \in \mathbb{A}_\Theta} \left\{ E\left[\left\langle \mathbf{Q}_{A_\theta}\ ,\ (\mathbf{C} \odot \mathbf{P}) \right\rangle_F \right] \right\} \\ &\approx\ \arg \min_{A_\theta \in \mathbb{A}_\Theta} \left\{ \left\langle\ \widehat{\mathbf{Q}_{A_\theta}}\ ,\ \mathbf{C} \odot \mathbf{E}\left[\mathbf{P}\right] \right\rangle_F \right\} \end{aligned} \tag{17}$$

As a result of Assumptions 1, 2, 3, 4, and 5 from Chapter I, expected values for elements of the cost, prevalence, and conditional probability matrices may all be analyzed and estimated independently of elements of any other matrix appearing in (17), prior to using them in calculation of Bayes' risk when employing the RRF.

We now consider the effect on $\mathbf{E}[\mathbf{P}]$ of varying our assumptions on $\mathbf{P}$. These assumptions may take many different forms. For example, we may simply consider that we already have an acceptable estimate of $\mathbf{P}$, and treat it as a constant, bringing us back to a form like that of Equation (10), Chapter II. We may also populate its rows with the transpose mean vector of a joint statistical distribution, such as a joint uniform distribution representing no knowledge of prior probabilities, or some other jointly continuous fixed-support probability distribution function, such as a multivariate Beta distribution. Finally, we may simply impose a joint weighting based on expert knowledge and belief (a.k.a., a *belief function*, which is actually a more general type of weighting than a probability distribution function, with potentially greater utility for actual end-users of classification systems [28, pp. 38-39]. In the latter case, we do not end up with a classical risk, but rather a *fuzzy* risk. Since the case where all random variables in the prevalence matrix $\mathbf{P}$ are constants is a matter of simple algebra, we shall examine a small sampling of more interesting possibilities.

3.2 Completely Unknown Class Prevalences

As noted in Chapter I, for an n-class classification system, exactly (n − 1) of the class prevalences are distributed over a standard (n − 1)-simplex, and the remaining class prevalence is found by solving the conjunctive equation inherent in each row of the stochastic matrix $\mathbf{P}$ of class prevalences. Observing Theorem 2, Appendix A:

$$\mathrm{m}! \;=\; \frac{1}{\int_0^1 \int_0^{1-\mathrm{p}_1} \int_0^{1-\mathrm{p}_1-\mathrm{p}_2} \ldots \int_0^{1-\sum_{\mathrm{j}=1}^{\mathrm{m}-1}(\mathrm{p}_\mathrm{j})} \; d\mathrm{p}_{\,\mathrm{m}} \ldots d\mathrm{p}_3\, d\mathrm{p}_2\, d\mathrm{p}_1} \tag{18}$$

we see the integral in the denominator of (18) is simply over the standard m-simplex $\Delta_\mathrm{m} \subset [0,1]^\mathrm{m}$ (see Definition 23, Chapter I). Apply this observation and (18) to the standard simplex $\Delta_{\mathrm{n}-1}$ from which the first $(\mathrm{n}-1)$ class prevalences are drawn, yielding:

$$\begin{aligned} \int_{\Delta_{\mathrm{n}-1}} d\mathrm{p}_{\,\mathrm{n}-1} \ldots d\mathrm{p}_1 \;&=\; \int_0^1 \int_0^{1-\mathrm{p}_1} \int_0^{1-\mathrm{p}_1-\mathrm{p}_2} \ldots \int_0^{1-\sum_{\mathrm{j}=1}^{\mathrm{n}-2}(\mathrm{p}_\mathrm{j})} d\mathrm{p}_{\,\mathrm{n}-1} \ldots d\mathrm{p}_3\, d\mathrm{p}_2\, d\mathrm{p}_1 \\ &=\; \frac{1}{(\mathrm{n}-1)!} \end{aligned} \tag{19}$$

Assuming nothing whatsoever is known about the prior probabilities, a jointly continuous uniform probability density function $W_{\mathbf{P},\mathrm{uniform}}(\mathrm{p}_1, \ldots, \mathrm{p}_{\,\mathrm{n}-1})$ of $(\mathrm{n}-1)$ class prevalences over the standard $(\mathrm{n}-1)$-simplex, satisfying $\int_{\Delta_{\mathrm{n}-1}} W_{\mathbf{P},\mathrm{uniform}}\big(\mathrm{p}_1, \ldots, \mathrm{p}_{\,\mathrm{n}-1}\big)\; d\mathrm{p}_{\,\mathrm{n}-1} \ldots d\mathrm{p}_1 \;=\; 1$ is then given by [3, p. 568]:

$$W_{\mathbf{P},\mathrm{uniform}}\big(\mathrm{p}_1, \ldots, \mathrm{p}_{\,\mathrm{n}-1}\big) \;=\; \begin{cases} (\mathrm{n}-1)!\,, & \big(\mathrm{p}_1, \ldots, \mathrm{p}_{\,\mathrm{n}-1}\big) \;\in\; \Delta_{\mathrm{n}-1} \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

If we consider the quantity $\mathbf{E}\,[\mathbf{P}]$ appearing in (17) above to be the *matrix* of expected values whose typical element is:

$$\left[\int_{\Delta_{\mathrm{n}-1}} \mathrm{p}_\mathrm{ij}\, W_{\mathbf{P}_\mathrm{j},\mathrm{uniform}}\big(\mathrm{p}_{\,1\mathrm{j}}, \ldots, \mathrm{p}_{\,\mathrm{n}-1,\mathrm{j}}\big)\; d\mathrm{p}_{\,\mathrm{n}-1,\mathrm{j}} \ldots d\mathrm{p}_{\,1\mathrm{j}} \right]_\mathrm{ij}$$

we may simplify our evaluation of these integrals by recalling that each row is identical and that all entries in a given row i, save $\mathrm{p_{in}} = \left(1 - \sum_{\mathrm{j}=1}^{\mathrm{n}-1} \mathrm{p_{ij}}\right)$, are the class prevalences $\left\{\mathrm{p_{ij}}\right\}_{\mathrm{j}=1}^{\mathrm{n}-1}$, for all rows $\mathrm{i} = 1, 2, 3 \ldots, \mathrm{n}$. Also, since the rows are identical, there is no need to keep the subscript i when referring to a prior probability $\mathrm{p_j}$ for class $\mathcal{E}_{\mathrm{j}}$.

It is *crucial* to state here that even though we use the words *first* and *last* to describe the class prevalences, the order in which the so-called *first* $(\mathrm{n}-1)$ class prevalences are randomly drawn from their joint distribution has nothing to do with the ordering of the index set Λ by which we link them to elements of the label set.

Since $\mathrm{p_n} = 1 - \sum_{\mathrm{j}=1}^{\mathrm{n}-1} \mathrm{p_j}$ is a function of the $(\mathrm{n}-1)$ class prevalences whose joint distribution is $W_{\mathbf{P},\mathrm{uniform}}(\mathrm{p}_1, \ldots, \mathrm{p}_{\mathrm{n}-1})$, we may use the weighting in (20) to calculate an expected value $E\left(\mathrm{p}_{\mathrm{n}}\right)$, using some of the equation patterns seen in the proof of Theorem 2, Appendix A:

$$
\begin{aligned}
E\left(\mathrm{p_n}\right) &= E\left(1 - \sum_{\mathrm{j}=1}^{\mathrm{n}-1} \mathrm{p_j}\right) \\
&= \int_{\Delta_{\mathrm{n}-1}} \left(1 - \sum_{\mathrm{j}=1}^{\mathrm{n}-1} \mathrm{p_j}\right) W_{\mathbf{P}}\left(\mathrm{p}_1 \ldots \mathrm{p}_{\mathrm{n}-1}\right)\, d\mathrm{p}_{\mathrm{n}-1}, \ldots, d\mathrm{p}_1 \\
&= \int_{\Delta_{\mathrm{n}-1}} \left(1 - \sum_{\mathrm{j}=1}^{\mathrm{n}-1} \mathrm{p_j}\right) (\mathrm{n}-1)!\, d\mathrm{p}_{\mathrm{n}-1} \ldots d\mathrm{p}_1 \\
&= (\mathrm{n}-1)! \int_{\Delta_{\mathrm{n}-1}} \left(1 - \sum_{\mathrm{j}=1}^{\mathrm{n}-1} \mathrm{p_j}\right) d\mathrm{p}_{\mathrm{n}-1} \ldots d\mathrm{p}_1 \qquad (21) \\
&= (\mathrm{n}-1)! \int_0^1 \int_0^{1-\mathrm{p}_1} \ldots \int_0^{1-\sum_{\mathrm{j}=1}^{\mathrm{n}-2} (\mathrm{p_j})} \left(1 - \sum_{\mathrm{j}=1}^{\mathrm{n}-1} \mathrm{p_j}\right) d\mathrm{p}_{\mathrm{n}-1} \ldots d\mathrm{p}_2\, d\mathrm{p}_1 \\
&= (\mathrm{n}-1)! \left(\frac{1}{\mathrm{n}!}\right), \text{ by (28) and (29), Theorem 2 proof, Appendix A} \\
&= \frac{1}{\mathrm{n}}
\end{aligned}
$$

We may calculate expected values for the other (n − 1) class prevalences in a row by means of Corollary 1, Appendix A, which is known to be true for positive integers less than 48 (i.e., for most practical classification purposes):

$$\begin{aligned}(\mathrm{m}+1)! \;&=\; \frac{1}{\int_0^1 \int_0^{1-\mathrm{p}_1} \int_0^{1-\mathrm{p}_1-\mathrm{p}_2} \ldots \int_0^{\sum_{\mathrm{i}=1}^{\mathrm{m}-1}(\mathrm{p}_\mathrm{i})} \; \mathrm{p}_\mathrm{j}\; d\mathrm{p}_\mathrm{m} \ldots d\mathrm{p}_3\, d\mathrm{p}_2\, d\mathrm{p}_1} \\ &\equiv\; \frac{1}{\int_{\Delta_\mathrm{m}} \mathrm{p}_\mathrm{j}\, \mathbf{dp}}, \quad \forall \;\; \mathrm{j} \;=\; 1,2,3,\; \ldots\; ,\mathrm{m}\end{aligned} \tag{22}$$

Apply (22) to each of the expected values $E\big(\mathrm{p}_\mathrm{j}\big)$, $\mathrm{j} \;=\; 1, \;\ldots\; , \mathrm{n}-1$:

$$\begin{aligned}E\big(\mathrm{p}_\mathrm{j}\big) \;&=\; \int_{\Delta_{\mathrm{n}-1}} \mathrm{p}_\mathrm{j}\, W_{\mathbf{P}}\big(\mathrm{p}_1, \;\ldots\; , \mathrm{p}_{\mathrm{n}-1}\big)\; d\mathrm{p}_{\mathrm{n}-1} \;\ldots\; d\mathrm{p}_1 \\ &=\; \int_{\Delta_{\mathrm{n}-1}} \mathrm{p}_\mathrm{j}\, (\mathrm{n}-1)!\, d\mathrm{p}_{\mathrm{n}-1} \;\ldots\; d\mathrm{p}_1 \\ &=\; (\mathrm{n}-1)! \int_{\Delta_{\mathrm{n}-1}} \mathrm{p}_\mathrm{j}\; d\mathrm{p}_{\mathrm{n}-1} \;\ldots\; d\mathrm{p}_1 \\ &=\; (\mathrm{n}-1)! \left(\frac{1}{\mathrm{n}!}\right) \\ &=\; \frac{1}{\mathrm{n}}, \quad \forall\, \mathrm{j} \;=\; 1, \;\ldots\; , \mathrm{n}-1\end{aligned} \tag{23}$$

so by (21), each entry in the $\mathrm{n} \times \mathrm{n}$ matrix $\mathbf{E}\big[\,\mathbf{P}_\mathrm{A}\,\big]$ is exactly $\frac{1}{\mathrm{n}}$; therefore, if we set all class prevalences equal to begin with, the resultant expected value matrix is the same as when we assume an underlying multivariate uniform distribution over $\Delta_{\mathrm{n}-1}$.

3.3 Limited Knowledge of Class Prevalences

Probability density functions, such as the multivariate uniform and Beta distributions, are only a specific kind of “weighting” function $W_{\mathbf{P}}$ to be used in evidential or probabilistic reasoning regarding the class prevalences, since the classical structure of probability is only a specific instance of an infinite number of ways to approach the so-called "doctrine of chances" [2], [28]. This is the reason we have chosen to denote the weighting function as $W_{\mathbf{P}}$ instead of the usual symbol $f_{\mathbf{P}}$ for a marginal probability density function of the class prevalences. Even though we leave the framework open for expansion, we will only consider one additional classical distribution—a multivariate general Beta.

The marginal versions of a general Beta distribution are very flexible and may even be made to approximate normal distributions over limited support intervals. The standard univariate Beta distribution has support on $[0,1]$, and thus has a set of two parameters (for shape), but the general form has four parameters, because it includes two parameters $\mathrm{S}_{\mathrm{lower}}$ and $\mathrm{S}_{\mathrm{upper}}$ giving the bounds of the support interval $[lower, upper] \subset \mathbb{R}$ over which it is defined [15, p. 210]. In this thesis we shall always define the lower support bound to be $lower\ =\ 0$, effectively reducing the number of parameters to three; further, we shall also consider only those marginal Beta distributions that have the potential to approximate a normal distribution with a mean of $\left(\frac{upper}{2}\right)$ over their support intervals; i.e., those symmetric about the midpoint of their support. This means the two shape parameters are equal, so we have reduced the total number of possibly unique parameters to two—one for shape, and one for support. We shall denote this special case of the general Beta distribution as $\beta(t,S)$, where t is the value of the common shape parameter, and S is the upper bound of the support interval $[0,S]$. In the case where a joint probability density function $W_{\mathbf{P},\beta}$ is defined over a standard simplex Δ_{m}, we shall indicate such joint support by the notation $\beta(\mathbf{t},\Delta_{\mathrm{m}})$, where $\mathbf{t} \in (0,\infty)^{\mathrm{m}}$ is a vector of the common shape parameters used in the symmetric marginal probability density functions.

The support parameter of the marginal distributions of the first $(\mathrm{n}-1)$ class prevalences randomly drawn over an $(\mathrm{n}-1)$-simplex is a function of all prevalences previously drawn. In fact, it is because of this that expected value calculations for any function $f\big(\mathrm{p}_1, \ldots, \mathrm{p}_{\mathrm{n}-1}\big)$ may then be performed by means of the operator:

$$\begin{aligned}\int_{\Delta_{\mathrm{n}-1}} (\,\cdot\,)\, W_{\mathbf{P}}\; d\mathrm{p}_{\mathrm{n}-1} \ldots d\mathrm{p}_1 \;&=\; \int_0^1 \ldots \int_0^{1-\sum_{\mathrm{j}=1}^{\mathrm{n}-2}(\mathrm{p}_\mathrm{j})} (\,\cdot\,)\, W_{\mathbf{P}}\big(\mathrm{p}_1, \ldots, \mathrm{p}_{\mathrm{n}-1}\big)\; d\mathrm{p}_{\mathrm{n}-1} \ldots d\mathrm{p}_1 \\ &=\; \int_0^1 W_{\mathbf{P},1} \ldots \int_0^{1-\sum_{\mathrm{j}=1}^{\mathrm{n}-2}(\mathrm{p}_\mathrm{j})} (\,\cdot\,) \left[\, W_{\mathbf{P},\mathrm{n}-1}\; d\mathrm{p}_{\mathrm{n}-1} \right] \ldots d\mathrm{p}_1\end{aligned} \tag{24}$$

by decomposing the joint distribution $W_{\mathbf{P}}\big(\mathrm{p}_1, \ldots, \mathrm{p}_{\mathrm{n}-1}\big)$ into a form allowing elimination of one variable at a time, working from the inside of the integral toward the outside:

$$W_{\mathbf{P}}\big(\mathrm{p}_1, \ldots, \mathrm{p}_{\mathrm{n}-1}\big) \;=\; \Big[W_{\mathbf{P},1}\big(\mathrm{p}_1\big)\Big]\, \Big[W_{\mathbf{P},2}\big(\mathrm{p}_1, \mathrm{p}_2\big)\Big] \ldots \Big[W_{\mathbf{P},\mathrm{n}-1}\big(\mathrm{p}_1, \ldots, \mathrm{p}_{\mathrm{n}-1}\big)\Big] \tag{25}$$

Considering a 3-class scenario, we attempt to approximate a bivariate normal distribution of the first two class prevalences randomly drawn from the standard 2-simplex. Figure 5 depicts a bivariate $\beta\big([5, 290], \Delta_2\big)$ probability distribution function $W_{\mathbf{P},\beta,(5,290),\Delta_2}\big(\mathrm{p}_1, \mathrm{p}_2\big)$ of two class prevalences over the standard 2-simplex. The values of the common shape parameters for the marginal probability density functions were chosen after examining Figures 6 and 7.

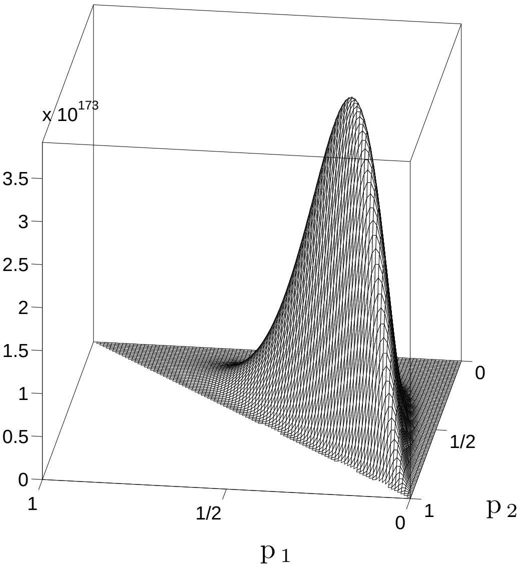

Figure 5: $W_{\mathbf{P},\beta,(5,290),\Delta_2}\left(\mathrm{p}_1,\mathrm{p}_2\right) \approx \frac{1.197\times 10^{178}\left(\mathrm{p}_1-\mathrm{p}_1^2\right)^4\left(\mathrm{p}_2-\mathrm{p}_1\mathrm{p}_2-\mathrm{p}_2^2\right)^{289}}{\left(1-\mathrm{p}_1\right)^{579}}$.

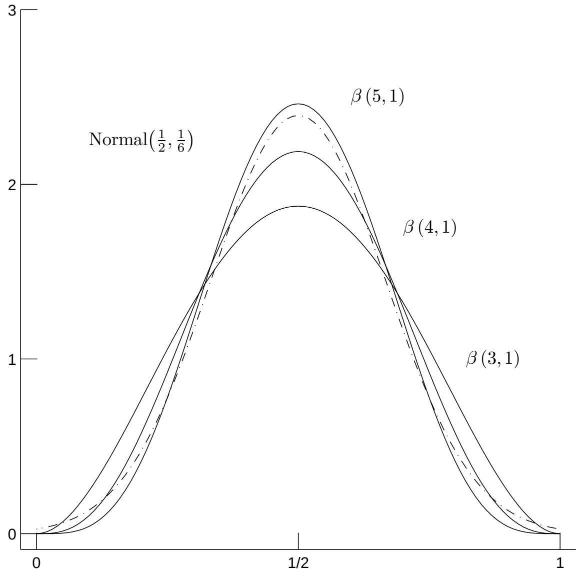

Figure 6: $\beta(5,1)$ Distribution as Normal$(\frac{1}{2},\frac{1}{6})$ Approximation.

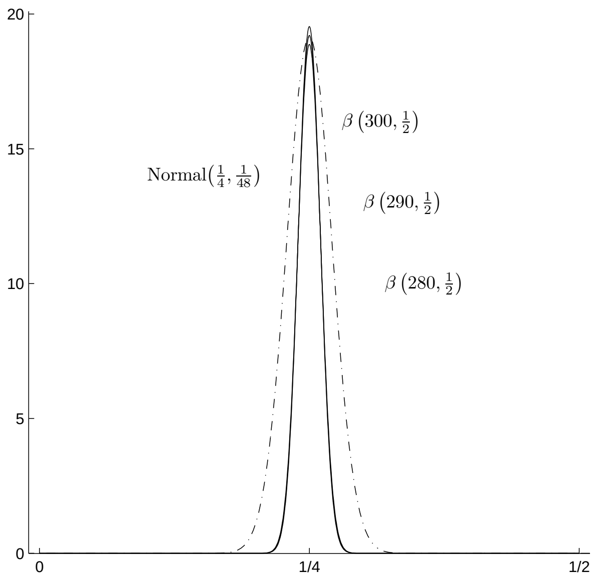

Figure 7: $\beta(290, \frac{1}{2})$ Distribution as Normal$(\frac{1}{4}, \frac{1}{48})$ Approximation.

IV. Application of Results to Actual Data

The Fisher Iris Data is a well-known data set consisting of four measurements (in millimeters) of various physical attributes for three subspecies of iris flowers—namely, Iris Setosa, Iris Versicolor, and Iris Virginica. There are 50 such sets of measurements per species, allowing for great flexibility when varying class distributions such that the data set is always of significant size. Principal components analysis (PCA) of the data reveals that the first two principal components account for about 98% of the variance in the data; therefore, we sped up computation by only using the component scores from these two components. The PCA scores for these two components are shown in Figure 8.

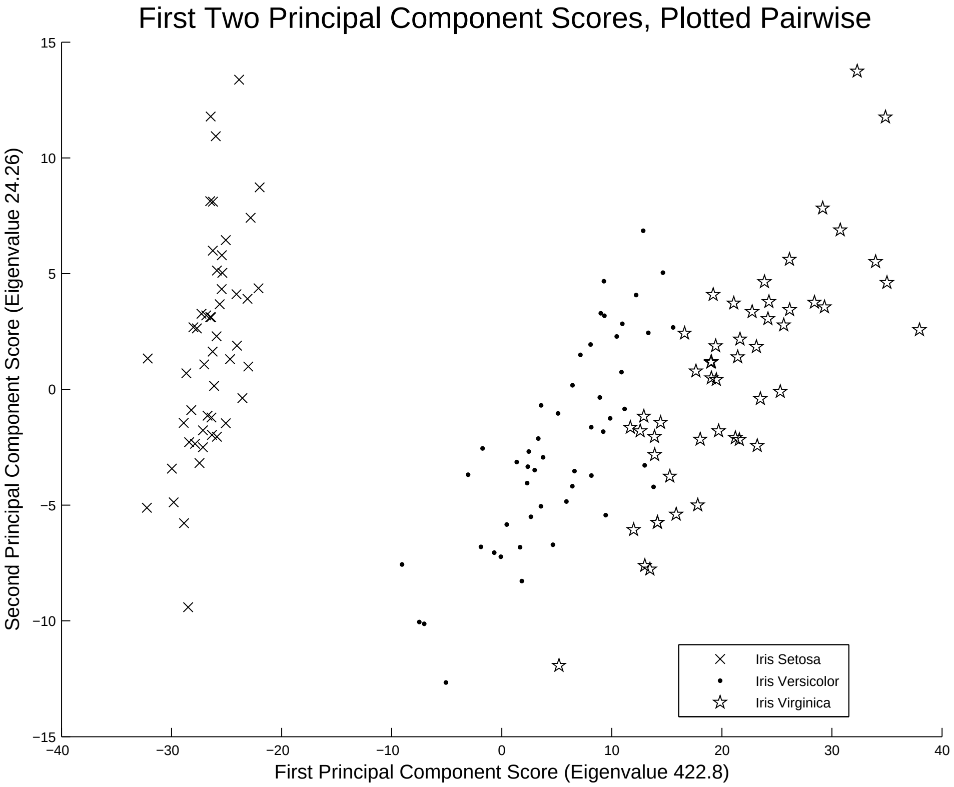

Figure 8: First Two Principal Component Scores, Fisher Iris Data.

We used Probabilistic Neural Net (PNN) classifiers trained with data distributed amongst the classes according to a set of three positive numbers summing to 1, the first two of which were drawn from a specific bivariate distribution assumed to exist over the standard 2-simplex, rounding determining the actual prevalence, which may not be exactly equal to the goal prevalence actually drawn due to the fact that one cannot classify a partial exemplar. We cast Iris Setosa as Class 1, Iris Versicolor as Class 2, Iris Virginica as Class 3, applying the uniform and symmetric Beta distributions examined in Chapter III, Sections 3.2 and 3.3 to the PCA score data to test the validity of Assumption 1 from Chapter I, while comparing the performance of the ROC Risk Functional (RRF) to Accuracy.

To test the validity of Assumption 1 from Chapter I, we used the non-parametric Kendall's Tau Correlation Coefficient statistical test, with a null hypothesis of no dependence between a class-conditional probability estimate $\widehat{\mathrm{q}_{\mathrm{i|j}}}$ and the prevalence $\mathrm{p_j}$ of the class $\mathcal{E}_{\mathrm{j}}$ upon which it is conditioned [11, pp. 404-405]. Note that we assumed it sufficient to test only between a conditional probability estimate $\widehat{\mathrm{q}_{\mathrm{i|j}}(A_\theta)}$ and the prevalence $\mathrm{p_j}$ of the class upon which it is conditioned, since the class prevalence $\mathrm{p_j}$ actually appears in the formulas for $\mathrm{q}_{\mathrm{i|j}}, \quad \forall\ \mathrm{i} = 1, \ldots, \mathrm{n}$ (see Definition 16, Chapter I). Under this assumption, nine separate Kendall's Tau tests were performed after each set of 37 replicates, testing for independence between each of the 9 sets of 37 class-conditional probabilities and a similar population of actual class prevalences for the class upon which they are conditioned, reporting the mean absolute value of Kendall's Tau Correlation Coefficients and corresponding mean p-values from those tests in a pair of 9×9 matrices.

A Monte Carlo simulated power analysis algorithm provided a 99% confidence interval of $(0.8083, 0.8282)$ for the power of a test with 37 sample points and an alternative hypothesis that the absolute value of Kendall's Tau Correlation Coefficient was as great as 0.4 and considering p-values of less than 0.15 statistically significant.

For both the uniform and beta scenarios, we trained a PNN classifier on the subsets of the data set derived by drawing the first two class prevalences randomly from the appropriate bivariate distribution until its randomly determined membership count was met, maximizing the overall size of the training data set according to the constraints of the randomly determined prevalences. We disallowed instances of zero class membership for any class, since Assumption 1 from Chapter I only applies to non-zero class prevalences. We validated the classifiers via the Lachenbruch holdout method, which yields a very precise estimate of Error, called the "Actual Error Rrate" (AER), where Error $\equiv$ $(1 - \text{Accuracy})$ [1], [17, p. 4]. In this method, the classifier is trained on all but one exemplar at a time and then that exemplar is classified to populate the contingency matrix. After the contingency matrix is completely populated in this way by repeating the Lachenbruch holdout procedure for each exemplar from a randomly chosen set of training data, the transpose stochastic confusion matrix was formed; then, the entire process listed above was repeated a total of 37 times, for both the uniform and beta scenarios.

A PNN uses a continuous threshold parameter called the spread. This is the common standard deviation of the small multivariate normal probability density functions that are constructed with each training exemplar as the mean vector, then summed and normalized to form the "Parzen Window" probability density functions for each class during training [5, pp. 164-166]. We standardized the data before training and validation, enabling us to vary the spread parameter for the PNN from 0.001 to 1.001 with confidence of not needing to go any larger with the spread [1]. If one sought to classify a new exemplar according to such a classifier, one would need to subtract the grand mean of the training data and divide by its standard deviation to obtain a standardized form of the exemplar.

We performed a “spread study” by taking 10 equally-spaced steps of 0.1 between 0.001 and 1.001, performing the 37 replications mentioned above at each point for both the uniform and beta scenarios. We found the value of the spread parameter such that a classification system based on that parameter minimized Bayes risk under a certain assumed cost regime and class prevalence distribution. The conditional probability matrix estimate used for a classification system based on a particular spread parameter and class prevalence distribution was the mean of the 37 confusion matrices produced by the experiment at that particular spread parameter, and for that particular distribution of class prevalences. We calculated Bayes risk under two different fixed-cost regimes for each of the prevalence distribution scenarios, but when the assumed distribution of prior probabilities was held constant between risk calculations for different cost regimes, we used the same conditional probability matrix estimate for both calculations. The two cost regimes used are shown in Tables 3 and 4.

Table 3: Cost Regime 1.

$$\begin{array}{r c} \text{Cost Regime 1} & \begin{array}{ccc} \text{Actual Class: 1} & \text{Actual Class: 2} & \text{Actual Class: 3} \end{array} \\ \begin{array}{r} \text{Labeled Class: 1} \\ \text{Labeled Class: 2} \\ \text{Labeled Class: 3} \end{array} & \left[\begin{array}{ccc} 0 & 5 & 5 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{array}\right] \end{array}$$

Table 4: Cost Regime 2.

$$\begin{array}{r c} \text{Cost Regime 2} & \begin{array}{ccc} \text{Actual Class: 1} & \text{Actual Class: 2} & \text{Actual Class: 3} \end{array} \\ \begin{array}{r} \text{Labeled Class: 1} \\ \text{Labeled Class: 2} \\ \text{Labeled Class: 3} \end{array} & \left[\begin{array}{ccc} 1 & 10 & 2 \\ 2 & 1 & 2 \\ 2 & 10 & 1 \end{array}\right] \end{array}$$

4.1 Uniform Distribution Scenario

As shown in Figures 9 and 10, spread parameter value $\theta = 0.301$ minimized Bayes Risk over two separate fixed cost scenarios. Figure 11 shows that this same value of the spread also minimized the AER.

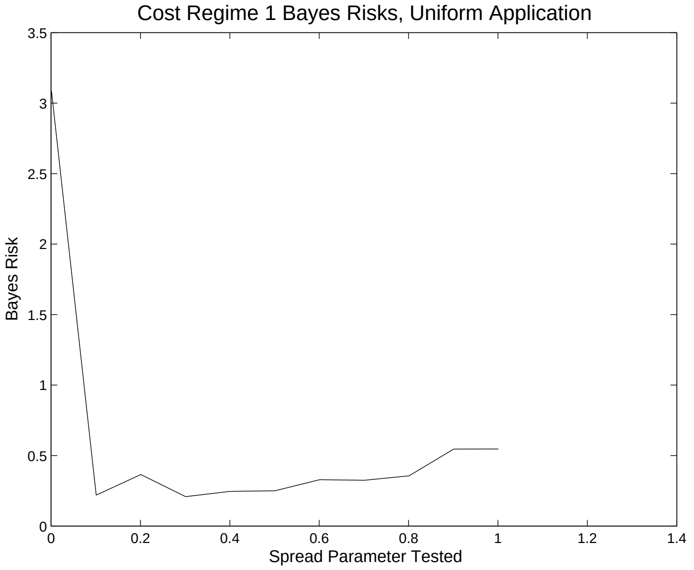

Figure 9: Bayes Risks for Cost Regime 1 , Uniform Application.

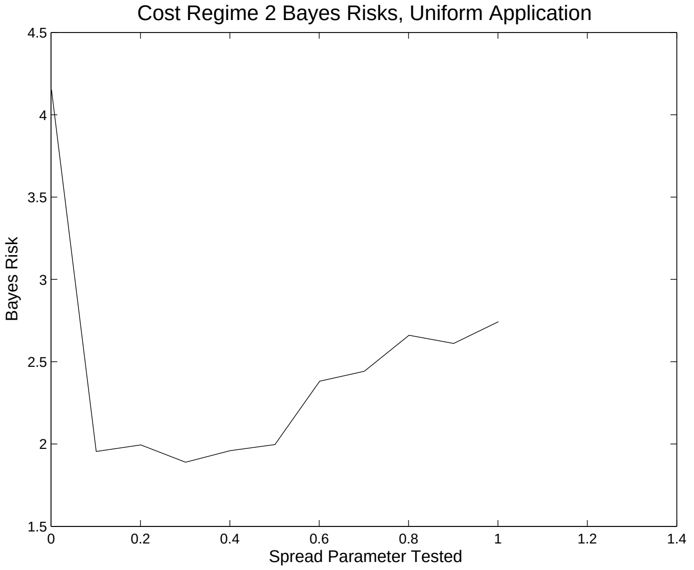

Figure 10: Bayes Risks for Cost Regime 2 , Uniform Application.

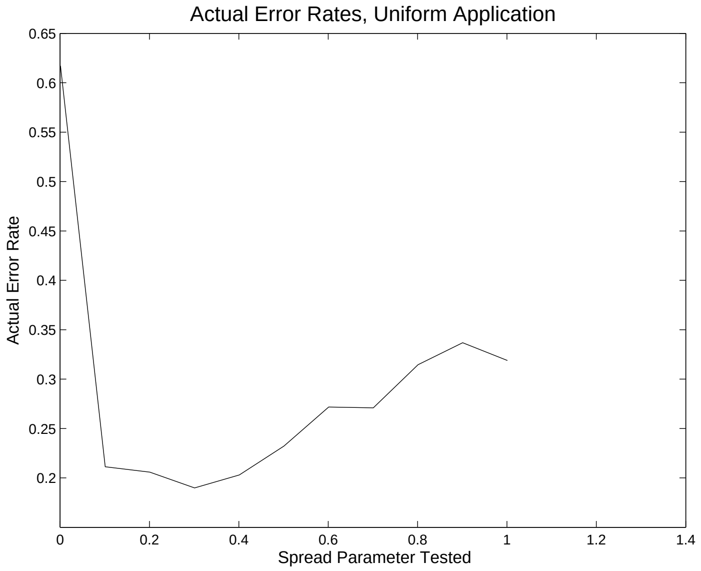

Figure 11: Actual Error Rates, Uniform Application.

As we can see from the mean p-values and absolute correlations in Table 5, Assumption 1 from Chapter I appears to have been violated in quite a few cases, especially in the lower right-hand corner of the table, corresponding to classification decisions involving the Iris Versicolor and Iris Virginica species. It should be noted that for the uniform scenario, lower p-values and higher correlations appeared in the areas of Table 5 corresponding to these species, no matter how we rearranged the order of which species were assigned to which class numbers. This may be related to the relative difficulty in distinguishing between these two species, as illustrated in Figure 8 above.

Finally, it is worth noting that Accuracy-based analysis, wherein the goal is to minimize the AER, yielded no different results than the RRF in this case.

Table 5: Mean Absolute Correlations and p-values, Uniform Application.

$$\begin{array}{cc}
\text{Correlations} & \begin{array}{ccc} \text{Actual Class: 1} & \text{Actual Class: 2} & \text{Actual Class: 3} \end{array} \\
\begin{array}{c} \text{Labeled Class: 1} \\ \text{Labeled Class: 2} \\ \text{Labeled Class: 3} \end{array} & \left[\begin{array}{ccc} 0.32 & 0.32 & 0.07 \\ 0.29 & 0.64 & 0.52 \\ 0.08 & 0.47 & 0.57 \end{array}\right]
\end{array}$$

$$\begin{array}{cc}
\text{p-values} & \begin{array}{ccc} \text{Actual Class: 1} & \text{Actual Class: 2} & \text{Actual Class: 3} \end{array} \\
\begin{array}{c} \text{Labeled Class: 1} \\ \text{Labeled Class: 2} \\ \text{Labeled Class: 3} \end{array} & \left[\begin{array}{ccc} 0.20 & 0.09 & 0.78 \\ 0.21 & 0.00 & 0.09 \\ 0.75 & 0.09 & 0.00 \end{array}\right]
\end{array}$$

4.2 Beta Distribution Scenario

As shown in Figures 12 and 13, spread parameter value $\theta = 0.201$ minimized Bayes Risk for Cost Regime 1, and a different spread parameter value $\theta = 0.401$ minimized Bayes Risk for Cost Regime 1. It is interesting to note, as displayed in Figure 14, that yet another spread parameter value $\theta = 0.301$, which was near the mean of the two parameters minimizing risk under the two cost scenarios, minimized the AER.

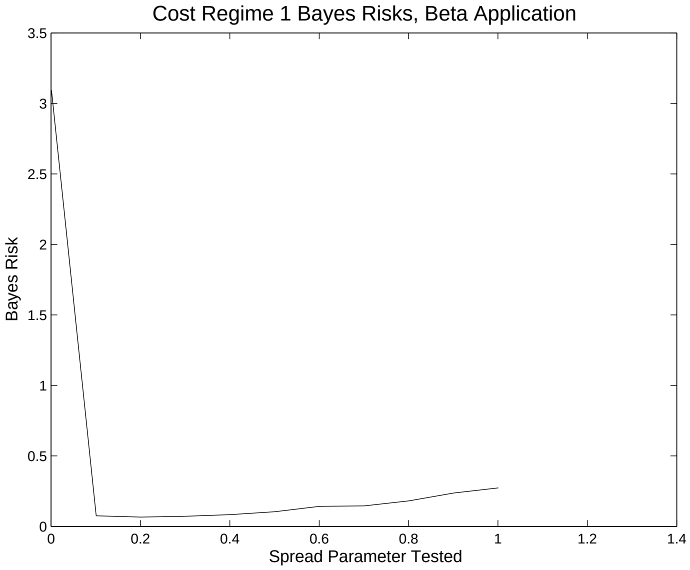

Figure 12: Bayes Risks for Cost Regime 1 , Beta Application.

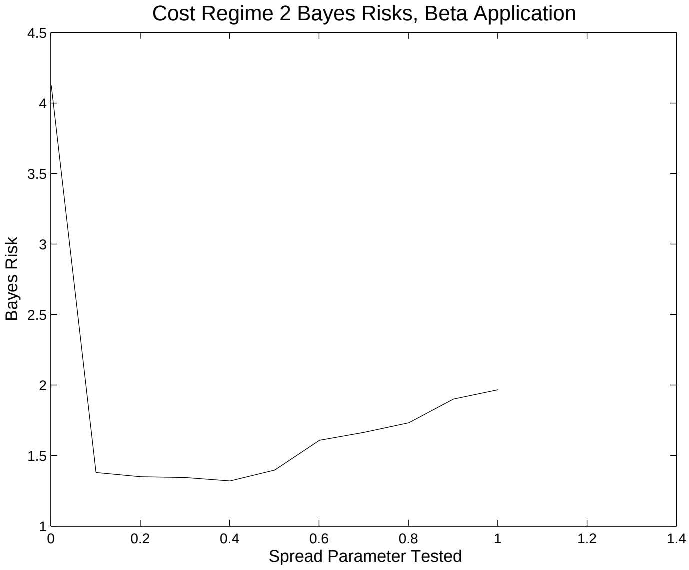

Figure 13: Bayes Risks for Cost Regime 2 , Beta Application.

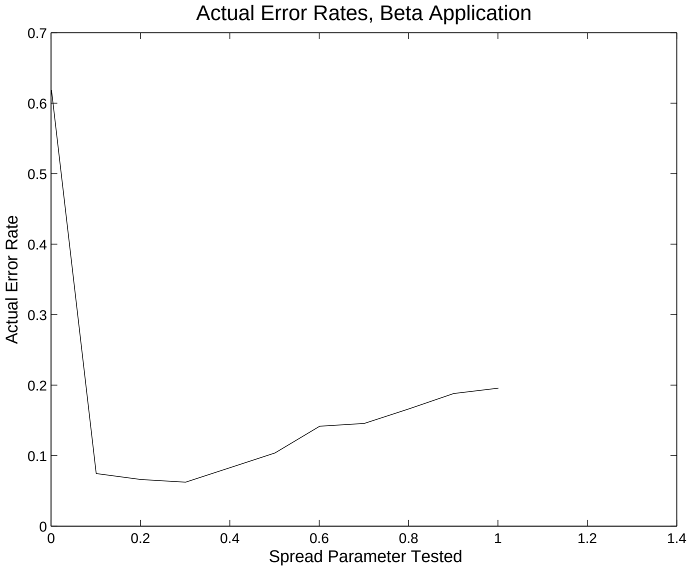

Figure 14: Actual Error Rates, Beta Application.

As we can see from the mean p-values and absolute correlations in Table 6, Assumption 1 from Chapter I appears to hold quite well in this scenario. It is interesting to note that the lowest correlations and highest p-values occurred in the $(1,3)$ and $(1,3)$ positions of Table 6. This may be related to the fact that, as shown in Figure 8 above, the

Classes 1 and 3, namely, Iris Setosa and Iris Virginica, are difficult to confuse.

Table 6: Mean Absolute Correlations and p-values, Beta Application.

$$\begin{array}{l|ccc|}
\text{Correlations} & \text{Actual Class: 1} & \text{Actual Class: 2} & \text{Actual Class: 3} \\
\text{Labeled Class: 1} & 0.09 & 0.16 & 0.04 \\
\text{Labeled Class: 2} & 0.09 & 0.26 & 0.19 \\
\text{Labeled Class: 3} & 0.00 & 0.19 & 0.23
\end{array}$$

$$\begin{array}{l|ccc|}
\text{p-values} & \text{Actual Class: 1} & \text{Actual Class: 2} & \text{Actual Class: 3} \\
\text{Labeled Class: 1} & 0.73 & 0.62 & 0.91 \\
\text{Labeled Class: 2} & 0.73 & 0.15 & 0.27 \\
\text{Labeled Class: 3} & 1.00 & 0.30 & 0.18
\end{array}$$

It should be noted here that if the manner of assigning class numbers to the species varied at all from the arrangement listed at the beginning of this Chapter, certain p-values began to be low and corresponding correlations high, similar to the case with the uniform application. This may be related to the fact that we had unequal means for the classes in the beta scenario, since the marginal probability distribution for the first class chosen always had mean $\left(\frac{1}{2}\right)$; thus, when a species that was not as easy as Iris Setosa to classify was chosen as Class 1, it tended to have a larger class prevalence, and thus more influence over the training process.

V. Conclusion and Research Suggestions

5.1 Summary of Application Results

With no knowledge of class prevalence distributions, persons training classifiers may wish to assume a uniform distribution. However, as has been shown, Assumption 1 from Chapter I may tend not to be met in this case, particularly if there are quite a few wrong decisions being made by the classifier. The Kendall's Tau test seems to be rather sensitive to these mistakes, and the fact that a confusion matrix should (in the ideal case) have only a few non-zero off-diagonal entries seems to create a situation with an excessive number of ties. The relationship between the relative number of times a classifier makes a mistake and does not seems to hold great power over these results; for example, there was one type of classification mistake that was *never* made in any case over all of the random trials performed, and so the correlation values for this element of the confusion matrix and the prevalence of the actual class over which the element was normalized was always exactly zero, with p-value 1. However, if just one mistake of a certain type occurred during classifier validation, this often resulted in a rather high correlation and a rather low p-value for the independence test for that particular class-conditional probability estimate.

Regardless of the fact that the Assumption 1 from Chapter I appears to have been violated, there are many cases in which the basic assumptions of an applied multivariate analysis technique may often be seriously violated, and yet the technique based upon these assumptions is still very useful and informative [1]. Therefore, I would recommend not eliminating risk-based comparison of classification systems until an investigation into such matters can be made, or another practical method of calculating risk is found that does not require such strict independence assumptions (if indeed such a practical method exists).

It would also appear that a more informative class prevalence distribution than the uniform tends to yield better results when testing Assumption 1 from Chapter I. The method for training classifiers, wherein training data is randomly chosen according to an assumed statistical or other distribution, may have applications for persons involved in the development of classification systems, because it eliminates human bias and allows the testing of said Assumption.

Based on the results of this thesis, I would advocate a paradigm shift toward risk-based comparison of classification systems in the field of ROC analysis, to allow both the users and producers of classification systems to have more confidence in the performance of these systems.

5.2 Suggestions for Further Research

Possible areas of further research are:

1. The field of belief functions may be more accessible to end-users as a potential weighting function on prior probabilities, since performing statistical experiments may be too expensive or difficult.
2. The framework of independence, if validated, leaves the door open for others to form and test risk over independently analyzed distributions of costs and class-conditional probabilities as well as class prevalences, if indeed such distributions may be found.
3. The feasibility of calculating the ROC convex hull (ROCCH) as a time-saving method for near-real time analysis of classification systems is still in question.

4. The possible need to test each class-conditional probability for independence against *all* class prevalences, not just the prevalence of the class upon which it is conditioned.

5. A better test statistic for independence, other than Kendall's Tau Correlation Coefficient, may exist.

6. An application for designed experiments to aid in spread studies or other such risk-based comparisons.

Appendix A. Mathematical Proofs

A.1 Conjecture Involving the Binomial Coefficients

Conjecture 1 (Relating to Binomial Coefficients)**.** *Given any positive integer* $\mathrm{m} \leq 47$*:*

$$\sum_{\mathrm{u}=0}^{\mathrm{m}} \left[\frac{(-1)^{\mathrm{u}}}{\mathrm{u}+2} \binom{\mathrm{m}}{\mathrm{u}} (\mathrm{m} + 2)(\mathrm{m} + 1) \right] = 1 \tag{26}$$

Proof. By exhaustion, directly calculated for $\mathrm{m} \leq 47$ using MATLAB® (calculation for integers greater than 47 exceeds machine precision limits, causing unavoidable computational error). □

A.2 Integrals Involving the Standard n*-Simplex*

Theorem 2 (Volume Under Standard Simplex)**.** *Given any positive integer* n*, along with any finite sequence* $\{\mathrm{x_i}\}_{\mathrm{i}=1}^{\mathrm{n}}$ *of real variables, the multiplicative inverse of the integral of the identity function over the standard* n*-simplex, as in Definition 23, Chapter I, is simply* $\mathrm{n}!$*:*

$$\frac{1}{\int_0^1 \int_0^{1-\mathrm{x}_1} \int_0^{1-\mathrm{x}_1-\mathrm{x}_2} \cdots \int_0^{1-\sum_{\mathrm{i}=1}^{\mathrm{n}-1}(\mathrm{x_i})} d\mathrm{x_n} \ldots d\mathrm{x}_3\, d\mathrm{x}_2\, d\mathrm{x}_1} = \mathrm{n}! \tag{27}$$

Proof. To prove the desired result, we shall prove its equivalent: that the value of the denominator on the left-hand side of (27) is $\left(\frac{1}{\mathrm{n}!}\right)$. We begin by performing the first three integrations indicated, working from the inside out, to determine if there is a consistent pattern:

$$
\begin{aligned}
&\int_0^1 \int_0^{1-x_1} \int_0^{1-x_1-x_2} \cdots \int_0^{1-\sum_{i=1}^{n-1}(x_i)} dx_n \,\ldots\, dx_3\, dx_2\, dx_1 \\
&\quad = \int_0^1 \int_0^{1-x_1} \cdots \int_0^{1-\sum_{i=1}^{n-2}(x_i)} \big[x_n\big]\Big|_{x_n=0}^{1-\sum_{i=1}^{n-1}(x_i)} dx_{n-1} \,\ldots\, dx_2\, dx_1 \\
&\quad * = \int_0^1 \int_0^{1-x_1} \cdots \int_0^{1-\sum_{i=1}^{n-2}(x_i)} \left(1-\sum_{i=1}^{n-1}\big[x_i\big]\right) dx_{n-1} \,\ldots\, dx_2\, dx_1 \\
&\quad = \int_0^1 \int_0^{1-x_1} \cdots \int_0^{1-\sum_{i=1}^{n-2}(x_i)} \left(1-\sum_{i=1}^{n-2}\big[x_i\big]-x_{n-1}\right) dx_{n-1} \,\ldots\, dx_2\, dx_1 \\
&\quad = \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{n-3}(x_i)} \left[-\left(\frac{1}{2}\right)\left(1-\sum_{i=1}^{n-2}[x_i]-x_{n-1}\right)^2\right]\Bigg|_{x_{n-1}=0}^{1-\sum_{i=1}^{n-2}(x_i)} dx_{n-2} \,\ldots\, dx_2\, dx_1 \\
&\quad * = \left(\frac{1}{2\,!}\right) \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{n-3}(x_i)} \left(1-\sum_{i=1}^{n-2}\big[x_i\big]\right)^2 dx_{n-2} \,\ldots\, dx_2\, dx_1 \\
&\quad = \left(\frac{1}{2\,!}\right) \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{n-3}(x_i)} \left(1-\sum_{i=1}^{n-3}\big[x_i\big]-x_{n-2}\right)^2 dx_{n-2} \,\ldots\, dx_2\, dx_1 \\
&\quad = \left(\frac{1}{2\,!}\right) \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{n-4}(x_i)} \left[-\left(\frac{1}{3}\right)\left(1-\sum_{i=1}^{n-3}[x_i]-x_{n-2}\right)^3\right]\Bigg|_{x_{n-2}=0}^{1-\sum_{i=1}^{n-3}(x_i)} dx_{n-3} \,\ldots\, dx_1 \\
&\quad * = \left(\frac{1}{3\,!}\right) \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{n-4}(x_i)} \left(1-\sum_{i=1}^{n-3}\big[x_i\big]\right)^3 dx_{n-3} \,\ldots\, dx_2\, dx_1
\end{aligned}
\tag{28}
$$

A consistent, predictable pattern is now recognizable on the lines denoted by an asterisk (*). When the largest remaining variable *index* is (n – k), there will be a constant $\left(\frac{1}{k\,!}\right)$ in front of the integral signs. Also, the resulting integrand will simply be the quantity $\left(1-\sum_{i=1}^{n-k}\big[x_i\big]\right)^k$, and the upper limit of integration for the innermost integral will be the quantity $\left(1-\sum_{i=1}^{n-(k+1)}\big[x_i\big]\right)$. We may now proceed to complete the calculation.

Performing repeated integration in this manner until the largest remaining variable index is (n – [n – 1] = 1), we obtain the desired result:

$$\begin{aligned}
&\int_0^1 \int_0^{1-x_1} \int_0^{1-x_1-x_2} \dots \int_0^{1-\sum_{i=1}^{n-1}(x_i)} dx_n \dots dx_3\, dx_2\, dx_1 \\
&= \left(\frac{1}{[n-1]\,!}\right) \int_0^{1-\sum_{i=1}^{n-[(n-1)+1]}(x_i)} \left(1 - \sum_{i=1}^{n-[n-1]} \left[x_i\right]\right)^{n-1} dx_{n-[n-1]} \\
&= \frac{1}{(n-1)\,!} \int_0^1 \left(1 - x_1\right)^{n-1} dx_1 \qquad (29) \\
&= \frac{1}{(n-1)\,!} \left[-\left(\frac{1}{n}\right) \left(1 - x_1\right)^{n} \right] \Bigg|_{x_1=0}^{1} \\
&= \frac{1}{n\,!}
\end{aligned}$$

□

Corollary 1 (Integral of an Axis Variable Over a Standard Simplex)**.** *Given any any finite sequence* $\left\{x_i\right\}_{i=1}^{n}$ *of axis variables for a standard* n*-simplex* Δ_n*, if* $\sum_{u=0}^{m}\left[\frac{(-1)^u}{u+2}\binom{m}{u}(m\,+\,2)\,(m\,+\,1)\right] = 1$ *holds true for all integers* $m \leq n$*, then the multiplicative inverse of the integral over* Δ_n *of any one of the axis variables* $x_j \in \left\{x_i\right\}_{i=1}^{n}$ *is* $(n+1)\,!$*:*

$$\frac{1}{\int_0^1 \int_0^{1-x_1} \int_0^{1-x_1-x_2} \dots \int_0^{1-\sum_{i=1}^{n-1}(x_i)} x_j \; dx_n \dots dx_3\, dx_2\, dx_1} = (n+1)\,!\,, \quad \forall\; j = 1, 2, 3\,, \dots\,, n \qquad (30)$$

Proof. To prove the desired result, we shall prove its equivalent: that the value of the denominator on the left-hand side of (30) is $\left(\frac{1}{[n+1]\,!}\right)$. Without loss of generality, consider the case where $j = k$, for some fixed $k = 1, 2, 3, \dots, n$:

$$
\begin{aligned}
&\int_0^1 \int_0^{1-x_1} \int_0^{1-x_1-x_2} \cdots \int_0^{1-\sum_{i=1}^{n-1}(x_i)} \left[x_k\right] \, , dx_n \, \ldots \, dx_3 \, dx_2 \, dx_1 \\
&\quad = \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-1}(x_i)} \int_0^{1-\sum_{i=1}^{k}(x_i)} \cdots \int_0^{1-\sum_{i=1}^{n-1}(x_i)} \left[x_k\right] \, , dx_n \, \ldots \, dx_{k+1} \, dx_k \, \ldots \, dx_1 \\
&\quad = \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-1}(x_i)} \int_0^{1-\sum_{i=1}^{k}(x_i)} \cdots \int_0^{1-\sum_{i=1}^{n-1}(x_i)} \left[x_k\right] \, , dx_n \, \ldots \, dx_{k+1} \, dx_k \, \ldots \, dx_1 \\
&\quad = \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-1}(x_i)} \left[x_k\right] \, , \left[\int_0^{1-\sum_{i=1}^{k}(x_i)} \cdots \int_0^{1-\sum_{i=1}^{n-1}(x_i)} dx_n \, \ldots \, dx_{k+1}\right] dx_k \, \ldots \, dx_1
\end{aligned}
\tag{31}
$$

We know from equation patterns in the proof of Theorem 2 above that the term in brackets on the last line of (31) is the integrand of the identity function integrated over Δ_n after $(n-k)$ integrals have been performed, working from the inside out, and that this integrand is simply the quantity $\left(\frac{1}{[n-k]\,!}\left[1-\sum_{i=1}^{k}(x_i)\right]^{n-k}\right)$; therefore, we may write:

$$
\begin{aligned}
&\int_0^1 \int_0^{1-x_1} \int_0^{1-x_1-x_2} \cdots \int_0^{1-\sum_{i=1}^{n-1}(x_i)} \left[x_k\right] \, , dx_n \, \ldots \, dx_3 \, dx_2 \, dx_1 \\
&\quad = \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-1}(x_i)} \left[x_k\right] \, , \left[\int_0^{1-\sum_{i=1}^{k}(x_i)} \cdots \int_0^{1-\sum_{i=1}^{n-1}(x_i)} dx_n \, \ldots \, dx_{k+1}\right] dx_k \, \ldots \, dx_1 \\
&\quad = \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-1}(x_i)} \left[x_k\right] \, , \left[\frac{1}{(n-k)\,!}\left(1-\sum_{i=1}^{k}\left[x_i\right]\right)^{n-k}\right] dx_k \, \ldots \, dx_1 \\
&\quad = \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-1}(x_i)} \left[x_k\right] \, , \left[\frac{1}{(n-k)\,!}\left(1-\sum_{i=1}^{k-1}\left[x_i\right]-x_k\right)^{n-k}\right] dx_k \, \ldots \, dx_1 \\
&\quad = \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-1}(x_i)} \left[x_k\right] \, , \left[\frac{1}{(n-k)\,!}\left(\left[1-\sum_{i=1}^{k-1}(x_i)\right]+\left[-x_k\right]\right)^{n-k}\right] dx_k \, \ldots \, dx_1
\end{aligned}
\tag{32}
$$

The Binomial Theorem states that $(a + b)^t = \sum_{u=0}^{t} \binom{t}{u} a^{t-u} b^u$, where $\binom{t}{u} \equiv \frac{t\,!}{(t-u)\,!\,u\,!}$. Applying this to the term in parentheses in (32), we may write:

$$\left(\left[1-\sum_{i=1}^{k-1}(x_i)\right]+\left[-x_k\right]\right)^{n-k} = \sum_{u=0}^{n-k}[-1]^u \binom{[n-k]}{u} \left[1-\sum_{i=1}^{k-1}(x_i)\right]^{(n-k)-u} \left[x_k\right]^u$$

Since we have arbitrarily fixed k, let us temporarily denote $m \equiv n - k$ to ease notational burdens. This enables us to rewrite (32) above as:

$$\begin{aligned}
&\int_0^1 \int_0^{1-x_1} \int_0^{1-x_1-x_2} \cdots \int_0^{1-\sum_{i=1}^{n-1}(x_i)} \left[x_k\right] \, , dx_n \ldots dx_3\, dx_2\, dx_1 \\
&= \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-1}(x_i)} \left[x_k\right] \, , \left[\frac{1}{(n-k)\,!}\left(\left[1-\sum_{i=1}^{k-1}(x_i)\right]+\left[-x_k\right]\right)^{n-k}\right] dx_k \ldots dx_1 \\
&\equiv \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-1}(x_i)} \left[x_k\right] \, , \left[\frac{1}{m\,!}\left(\left[1-\sum_{i=1}^{k-1}(x_i)\right]+\left[-x_k\right]\right)^{m}\right] dx_k \ldots dx_1 \\
&= \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-1}(x_i)} \left[x_k\right] \left(\frac{1}{m\,!}\right)\left[\sum_{u=0}^{m}[-1]^u \binom{m}{u}\left[1-\sum_{i=1}^{k-1}(x_i)\right]^{m-u}\left[x_k\right]^u\right] dx_k \ldots dx_1 \\
&= \left(\frac{1}{m\,!}\right)\sum_{u=0}^{m}(-1)^u\binom{m}{u}\left[\int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-1}(x_i)} \left[x_k\right]\left(1-\sum_{i=1}^{k-1}\left[x_i\right]\right)^{m-u}\left(x_k\right)^u dx_k \ldots dx_1\right] \\
&= \left(\frac{1}{m\,!}\right)\sum_{u=0}^{m}(-1)^u\binom{m}{u}\left[\int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-1}(x_i)} \left(1-\sum_{i=1}^{k-1}\left[x_i\right]\right)^{m-u}\left(x_k\right)^{u+1} dx_k \ldots dx_1\right]
\end{aligned} \tag{33}$$

where we have interchanged integration with finite summation. It now befalls us to evaluate the term in brackets in (33) above. We will perform one integration first:

$$\begin{aligned}
&\int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-1}(x_i)} \left(1-\sum_{i=1}^{k-1}[x_i]\right)^{m-u} \left(x_k\right)^{u+1} dx_k \ldots dx_1 \\
&\quad = \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-2}(x_i)} \left(1-\sum_{i=1}^{k-1}[x_i]\right)^{m-u} \left[\int_0^{1-\sum_{i=1}^{k-1}(x_i)} \left(x_k\right)^{u+1} dx_k\right] dx_{k-1} \ldots dx_1 \\
&\quad = \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-2}(x_i)} \left(1-\sum_{i=1}^{k-1}[x_i]\right)^{m-u} \left[\left(\frac{1}{u+2}\right)\left(x_k\right)^{u+2}\right]\Bigg|_{x_k=0}^{1-\sum_{i=1}^{k-1}(x_i)} dx_{k-1} \ldots dx_1 \\
&\quad = \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-2}(x_i)} \left(1-\sum_{i=1}^{k-1}[x_i]\right)^{m-u} \left[\left(\frac{1}{u+2}\right)\left(1-\sum_{i=1}^{k-1}[x_i]\right)^{u+2}\right] dx_{k-1} \ldots dx_1 \\
&\quad = \left(\frac{1}{u+2}\right)\int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-2}(x_i)} \left(1-\sum_{i=1}^{k-1}[x_i]\right)^{m-u} \left(1-\sum_{i=1}^{k-1}[x_i]\right)^{u+2} dx_{k-1} \ldots dx_1 \\
&\quad = \left(\frac{1}{u+2}\right)\int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-2}(x_i)} \left(1-\sum_{i=1}^{k-1}[x_i]\right)^{m+2} dx_{k-1} \ldots dx_1
\end{aligned}$$

Multiplying the bracketed term in (33) by the quantity $(u+2)$, we may now write:

$$\begin{aligned}
&(u+2)\int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-1}(x_i)} \left(1-\sum_{i=1}^{k-1}[x_i]\right)^{m-u} \left(x_k\right)^{u+1} dx_k \ldots dx_1 \\
&\quad = \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-2}(x_i)} \left(1-\sum_{i=1}^{k-1}[x_i]\right)^{m+2} dx_{k-1} \ldots dx_1 \\
&\quad = \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-2}(x_i)} \left(1-\sum_{i=1}^{k-2}[x_i]-x_{k-1}\right)^{m+2} dx_{k-1} \ldots dx_1 \\
&\quad = \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-3}(x_i)} \left[-\left(\frac{1}{m+3}\right)\left(1-\sum_{i=1}^{k-2}[x_i]-x_{k-1}\right)^{m+3}\right]\Bigg|_{x_{k-1}=0}^{1-\sum_{i=1}^{k-2}(x_i)} dx_{k-2} \ldots dx_1 \\
&\quad * = \left(\frac{1}{m+3}\right)\int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-3}(x_i)} \left(1-\sum_{i=1}^{k-2}[x_i]\right)^{m+3} dx_{k-2} \ldots dx_1 \\
&\quad = \left(\frac{1}{m+3}\right)\int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-3}(x_i)} \left(1-\sum_{i=1}^{k-3}[x_i]-x_{k-2}\right)^{m+3} dx_{k-2} \ldots dx_1
\end{aligned} \tag{34}$$

and if we additionally multiply the bracketed term in (33) by (m + 3), we may then write:

$$
\begin{aligned}
&(\mathrm{u}+2)(\mathrm{m}+3)\int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-1}(x_i)} \left(1-\sum_{i=1}^{k-1}\left[x_i\right]\right)^{m-u} \left(x_k\right)^{u+1} dx_k \ldots dx_1 \\
&= \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-3}(x_i)} \left(1-\sum_{i=1}^{k-3}\left[x_i\right]-x_{k-2}\right)^{m+3} dx_{k-2} \ldots dx_1 \\
&= \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-4}(x_i)} \left[-\left(\frac{1}{m+4}\right)\left(1-\sum_{i=1}^{k-3}\left[x_i\right]-x_{k-2}\right)^{m+4}\right]\Bigg|_{x_{k-2}=0}^{1-\sum_{i=1}^{k-3}(x_i)} dx_{k-3} \ldots dx_1 \\
* &= \left(\frac{1}{m+4}\right)\int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-4}(x_i)} \left(1-\sum_{i=1}^{k-3}\left[x_i\right]\right)^{m+4} dx_{k-3} \ldots dx_1
\end{aligned}
\tag{35}
$$

A consistent, predictable pattern is now recognizable on the lines denoted by an asterisk (*) in (34) and (35) above. When the largest remaining variable *index* is (k − j), there will appear a constant in front of the integral signs in the bracketed term in (33):

$$
\begin{aligned}
\left(\frac{1}{u+2}\right)\left(\frac{1}{m+3}\right)\left(\frac{1}{m+4}\right)\cdots\left(\frac{1}{m+[j+1]}\right) &= \left(\frac{1}{u+2}\right)\left[\frac{1}{\frac{(m+[j+1])\,!}{(m+2)\,!}}\right] \\
&= \left(\frac{1}{u+2}\right)\left[\frac{(m+2)\,!}{(m+[j+1])\,!}\right]
\end{aligned}
$$

Also, the resulting integrand will simply be the quantity $\left[\left(1-\sum_{i=1}^{k-j}\left[x_i\right]\right)^{m+[j+1]}\right]$, and the upper limit of integration for the innermost integral will be the quantity $\left(1-\sum_{i=1}^{k-[j+1]}\left[x_i\right]\right)$. We may now proceed to complete the calculation.

Performing repeated integration in this manner until the largest remaining variable index is $(k-[k-2] = 2)$, if we multiply the bracketed term in (33) by the quantity $(u+2)\left[\frac{(m+[(k-2)+1])!}{(m+2)!}\right] = (u+2)\left[\frac{(m+[k-1])!}{(m+2)!}\right] \equiv (u+2)\left[\frac{(n-1)!}{(m+2)!}\right]$, we may write:

$$
\begin{aligned}
(u+2)\left[\frac{(n-1)!}{(m+2)!}\right] &\int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-1}(x_i)} \left(1-\sum_{i=1}^{k-1}\big[x_i\big]\right)^{m-u} \big(x_k\big)^{u+1}\, dx_k \,\dots\, dx_1 \\
&= \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-[(k-2)+1]}(x_i)} \left(1-\sum_{i=1}^{k-[k-2]}\big[x_i\big]\right)^{m+[(k-2)+1]} dx_{k-[k-2]} \,\dots\, dx_1 \\
&= \int_0^1 \int_0^{1-\sum_{i=1}^{k-[k-1]}(x_i)} \left(1-\sum_{i=1}^{2}\big[x_i\big]\right)^{m+[k-1]} dx_2\, dx_1 \\
&\equiv \int_0^1 \int_0^{1-x_1} \big(1-x_1-x_2\big)^{n-1}\, dx_2\, dx_1
\end{aligned}
$$

i.e., the bracketed term in (33) may be simply written as:

$$
\begin{aligned}
\int_0^1 \int_0^{1-x_1} \int_0^{1-x_1-x_2} \cdots \int_0^{1-\sum_{i=1}^{n-1}(x_i)} & \big[x_k\big]\, dx_n \,\dots\, dx_3\, dx_2\, dx_1 \\
&= \left(\frac{1}{u+2}\right)\left[\frac{(m+2)!}{(n-1)!}\right] \int_0^1 \int_0^{1-x_1} \big(1-x_1-x_2\big)^{n-1} dx_2\, dx_1 \\
&= \left(\frac{1}{u+2}\right)\left[\frac{(m+2)!}{(n-1)!}\right] \int_0^1 \left[-\left(\frac{1}{n}\right)\big(1-x_1-x_2\big)^{n}\right]\Bigg|_{x_2=0}^{1-x_1} dx_1 \\
&= \left(\frac{1}{u+2}\right)\left[\frac{(m+2)!}{(n-1)!}\right] \left(\frac{1}{n}\right) \int_0^1 \big(1-x_1\big)^{n} dx_1 \\
&= \left[\frac{(m+2)!}{u+2}\right]\left(\frac{1}{n!}\right)\left[-\left(\frac{1}{n+1}\right)\big(1-x_1\big)^{n+1}\right]\Bigg|_{x_1=0}^{1} \\
&= \frac{(m+2)!}{u+2}\;\frac{1}{(n+1)!}
\end{aligned}
\tag{36}
$$

Substituting the result of (36) for the bracketed term in (33), the expression we wish to prove equal to $\left(\frac{1}{[n+1]!}\right)$ for the case where $j = k$ may now be written as:

$$
\begin{aligned}
&\int_0^1 \int_0^{1-x_1} \int_0^{1-x_1-x_2} \dots \int_0^{1-\sum_{i=1}^{n-1}(x_i)} \left[x_k\right]\, dx_n \dots dx_3\, dx_2\, dx_1 \\
&\quad = \left(\frac{1}{m\,!}\right) \sum_{u=0}^{m} (-1)^u \binom{m}{u} \left[\int_0^1 \dots \int_0^{1-\sum_{i=1}^{k-1}(x_i)} \left(1 - \sum_{i=1}^{k-1} \left[x_i\right]\right)^{m-u} \left(x_k\right)^{u+1} dx_k \dots dx_1 \right] \\
&\quad = \left(\frac{1}{m\,!}\right) \sum_{u=0}^{m} (-1)^u \binom{m}{u} \left[\frac{(m+2)\,!}{u+2} \; \frac{1}{(n+1)\,!} \right] \\
&\quad = \frac{1}{(n+1)\,!} \; \sum_{u=0}^{m} \left[\frac{(-1)^u}{u+2} \binom{m}{u} (m\,+\,2)\,(m\,+\,1) \right] \\
&\quad = \frac{1}{(n+1)\,!}
\end{aligned}
\tag{37}
$$

where the last step may be taken due to the fact that $m \equiv n - k$ for arbitrary $k = 1, \dots n$, hence $m \leq n$, and the result follows from the hypothesis.

□

Bibliography

1. K. W. Bauer. *Air Force Institute of Technology Class Lectures, OPER 685 and OPER 785*, Fall 2007 and Winter 2008, respectively.

2. R. T. Cox. "Probability, Frequency, and Reasonable Expectation". *American Journal of Physics* 14:1–13, 1946.

3. L. Devroye. *Non-Uniform Random Variate Generation.* Springer-Verlag, New York, 1968.

4. S. Dreisetl. "Training Multiclass Classifiers by Maximizing the Volume Under the ROC Surface". *Proceedings of EUROCAST 2007 - The 11th International Conference on Computer Aided Systems Theory* (Las Palmas de Gran Canaria, Spain). 2007.

5. R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* John Wiley & Sons, Inc. New York, 2001.

6. J. P. Egan. *Signal Detection Theory and ROC Analysis.* Academic Press, New York, 1975.

7. T. Fawcett and P. A. Flach. "A Response to Webb and Ting's *On the Application of ROC Analysis to Predict Classification Performance Under Varying Class Distributions*". *Machine Learning* 58:33-38, 2005.

8. C. Ferri, J. Hernandez-Orallo, and M. A. Salido. "Volume Under the ROC Surface for Multi-class Problems". *Proceedings of ECML 2003 - The 14th European Conference on Machine Learning* (Cavtat-Dubrovnik, Croatia). 2003.

9. J. E. Fieldsend and R. M. Everson. "Formulation and comparison of multi-class ROC surfaces". *Proceedings of the 2nd ROC Analysis in Machine Learning Workshop, part of ICML 2005 - The 22nd International Conference on Machine Learning* (Cavtat-Dubrovnik, Croatia). 2003.

10. P. A. Flach. "The Geometry of ROC Space: *Understanding Machine Learning Metrics through ROC Isometrics*". Proceedings of ICML 2003: The 20th International Conference on Machine Learning (Washington DC). 2003.

11. J. D. Gibbons and S. Chakraborti. *Non-parametric Statistical Inference.* Marcel Dekker, Inc. New York, 4th ed. 2003.

12. D. J. Hand and R. J. Till. "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems". *Machine Learning* 45:171–186, 2001.

13. J. A. Hanley and B. J. McNeil. "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve". *Radiology* 143:29–36, 1982.

14. J. Huang and C. X. Ling. “Using AUC and Accuracy in Evaluating Learning Algorithms”. *IEEE Transactions on Knowledge & Data Engineering* 17:299–310, 2005.

15. N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, Vol. 2. John Wiley & Sons, Inc. New York, 1995.

16. A. N. Kolmogorov. *Foundations of the Theory of Probability.* Edited by N. Morrison. Chelsea Publishing Company, New York, 2nd ed. 1956.

17. P. A. Lachenbruch and M. R. Mickey. “Estimation of Error Rates in Discriminant Analysis”. *Technometrics* 10:1–10, 1968.

18. L. C. Ludeman. *Random Processes: Filtering, Estimation, and Detection.* John Wiley & Sons, Inc. Hoboken NJ, 2003.

19. C. E. Metz. “Basic Principles of ROC Analysis”. *Seminars in Nuclear Medicine* 8:283–298, 1978.

20. L. B. Milev and S. Gy. Rèvèsz. “Bernstein’s Inequality for Multivariate Polynomials on the Standard Simplex”. *Journal of Inequalities and Applications* 2005:145–163, 2005.

21. D. Mossman. “Three-way ROCs”. *Medical Decision Making* 19:78–89, 1999.

22. C. T. Nakas and C. T. Yiannoutsos. “Ordered multiple-class ROC analysis with continuous measurements”. *Statistics in Medicine* 23:3437–3449, 2004.

23. M. L. Puri and P. K. Sen. *Nonparametric Methods in Multivariate Analysis.* John Wiley & Sons, Inc. New York, 1971.

24. F. Provost and T. Fawcett. “Robust classification systems for imprecise environments”. *Proceedings of AAAI 1998 - The 15th National Conference on Artificial Intelligence* (Madison WI). 1998.

25. F. Provost, T. Fawcett, and R. Kohavi. “The Case Against Accuracy Estimation for Comparing Induction Algorithms”. *Proceedings of ICML 1998 - The 15th International Conference on Machine Learning* (Madison WI). 1998.

26. S. Rosset. “Model Selection via the AUC”. *Proceedings of ICML 2004 - The 21st International Conference on Machine Learning* (Banff, Alberta, Canada). 2004.

27. H. L. Royden. *Real Analysis.* Prentice Hall, Englewood Cliffs NJ, 3rd ed. 1988.

28. G. Shafer. *A mathematical theory of evidence.* Princeton University Press, Princeton NJ, 1976.

29. D. Stirzaker. *Probability and Random Variables: a beginner’s guide.* Cambridge University Press, The Edinburgh Building, Cambridge, UK, 2005.

30. S. N. Thorsen, *The Application of Category Theory and Analysis of Receiver Operating Characteristics to Information Fusion.* Doctoral Dissertation. Air Force Institute of Technology (AU), Wright-Patterson AFB OH, December 2005. (ADA450338) (3209742).

31. S. N. Thorsen and M. E. Oxley. "A description of competing fusion systems". *Information Fusion* 7:346–360, 2006.

32. S. N. Thorsen and M. E. Oxley. "Quantifying the Robustness of Classification Systems". *Proceedings of the 15th Signal Processing, Sensor Fusion, and Target Recognition Conference* (Kissimmee FL). 2006.

33. D. D. Wackerly, W. Mendenhall III, and R. L. Scheaffer. *Mathematical Statistics with Applications.* Duxbury, Pacific Grove CA, 6th ed. 2002.

34. W. Waegeman, B. De Baets, and L. Boullart. "ROC analysis in ordinal regression learning". *Pattern Recognition Letters* 29:1–9, 2008.

35. G. I. Webb and K. M. Ting. "On the Application of ROC Analysis to Predict Classification Performance Under Varying Class Distributions". *Machine Learning* 58:25-32, 2005.

36. S. M. Winkler, M. Affenzeller, and S. Wagner. "Sets of Receiver Operating Characteristic Curves and their Use in the Evaluation of Multi-Class Classification". *Proceedings of GECCO 2006 - The 8th annual conference on Genetic and Evolutionary Computation* (Bonn, Germany). 2005.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
March 2008

2. REPORT TYPE
Master's Thesis

3. DATES COVERED *(From - To)*
June 2007 - March 2008

4. TITLE AND SUBTITLE
Risk-Based Comparison of Classification Systems

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Wagenman, Seth B.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Air Force Institute of Technology, Graduate School of Engineering and Management
(AFIT/EN), 2950 Hobson Way, WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT/GAM/ENC/08-01

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
intentionally left blank

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Performance measures for families of classification system families that rely upon the analysis of receiver operating characteristics (ROCs), such as area under the ROC curve (AUC), often fail to fully address the issue of risk, especially for classification systems involving more than two classes. For the general case, we denote matrices of class prevalences, costs, and class-conditional probabilities, and assume costs are subjectively fixed, acceptable estimates for expected values of class-conditional probabilities exist, and mutual independence between a variable in one such matrix and those of any other matrix. The ROC Risk Functional (RRF), valid for any finite number of classes, has an associated parameter argument, that which specifies a member of a family of classification systems, and which system minimizes Bayes risk over the family. We typify joint distributions for class prevalences over standard simplices by means of uniform and beta distributions, and create a family of classification systems using actual data, testing independence assumptions under two such class prevalence distributions. We minimize risk under two different sets of costs.

15. SUBJECT TERMS
Classification, Risk, Costs, Risk Analysis, ROC Analysis, Classification Systems, Performance

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
UU	UU	UU	UU	75	Steven N. Thorsen
					19b. TELEPHONE NUMBER *(Include area code)* (937) 255-3636 x 4584