9-18-2014

Statistical Inference on Optimal Points to Evaluate Multi-State Classification Systems

Katherine A. Batterton

STATISTICAL INFERENCE ON OPTIMAL POINTS TO EVALUATE MULTI-STATE CLASSIFICATION SYSTEMS

DISSERTATION

Katherine Anne Batterton, Captain, USAF

AFIT–ENC–DS–14–S–02

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

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

AFIT–ENC–DS–14–S–02

STATISTICAL INFERENCE ON OPTIMAL POINTS TO EVALUATE MULTI-STATE CLASSIFICATION SYSTEMS

DISSERTATION

Presented to the Faculty

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 Doctor of Philosophy

Katherine Anne Batterton, B.S., M.S.

Captain, USAF

September 2014

AFIT–ENC–DS–14–S–02

STATISTICAL INFERENCE ON OPTIMAL POINTS TO EVALUATE MULTI-STATE CLASSIFICATION SYSTEMS

DISSERTATION

Katherine Anne Batterton, B.S., M.S.
Captain, USAF

Approved:

//SIGNED//	16 Jun 2014
Christine M. Schubert Kabban, Ph.D. (Chairman)	Date
//SIGNED//	27 May 2014
Lt Col Richard L. Warr, Ph.D. (Member)	Date
//SIGNED//	29 May 2014
Kenneth W. Bauer, Ph.D. (Member)	Date

Accepted:

//SIGNED//	25 Jun 2014
ADEDEJI B. BADIRU, Ph.D. Dean, Graduate School of Engineering and Management	Date

Abstract

In decision making, an optimal point represents the settings for which a classification system should be operated to achieve maximum performance. Clearly, these optimal points are of great importance in classification theory. Not only is the selection of the optimal point of interest, but quantifying the uncertainty in the optimal point and its performance is also important.

The Youden index is a metric currently employed for selection and performance quantification of optimal points for classification system families. The Youden index quantifies the correct classification rates of a classification system, and its confidence interval quantifies the uncertainty in this measurement. This metric currently focuses on two or three classes, and only allows for the utility of correct classifications and the cost of total misclassifications to be considered. An alternative to this metric for three or more classes is a cost function which considers the sum of incorrect classification rates. This new metric is preferable as it can include class prevalences and costs associated with every classification. In multi-class settings this informs better decisions and inferences on optimal points.

The work in this dissertation develops theory and methods for confidence intervals on a metric based on misclassification rates, Bayes Cost, and where possible, the thresholds found for an optimal point using Bayes Cost. Hypothesis tests for Bayes Cost are also developed to test a classification systems performance or compare systems with an emphasis on classification systems involving three or more classes. Performance of the newly proposed methods is demonstrated with simulation.

For my parents who provided me with the tools,
and for Dain who kept me sane enough to use them.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor Dr. Christine Schubert Kabban for all of her time and support. I could not have completed this work without her guidance and encouragement. Additionally, I would like to thank Dr. Richard Warr and Dr. Kenneth Bauer for serving on my committee and providing their time and knowledge which directly improved this dissertation. I would also like to express my appreciation to Dr. Richard Martin for his thoughtful feedback. Finally, I would like to thank the Department of Mathematical Sciences at the United States Air Force Academy and the department head, Col John Andrew, for giving me this opportunity.

Katherine Anne Batterton

Table of Contents

Page

Page

List of Figures

List of Tables

Table Page

List of Acronyms

Acronym	Definition
ADI	adipose
AN	asymptotic normal
BC	Bayes Cost
BCa	bias corrected and accelerated
BP	basic percentile
CAR	carcinoma
CDF	cumulative distribution function
CI	confidence interval
CLT	Central Limit Theorem
CON	connective
CSF	classification system family
FAD	fibro-adenoma
GCI	generalized confidence interval
GLA	glandular
GPQ	generalized pivotal quantity
LRT	likelihood ratio test
MAS	mastopathy
MLE	maximum likelihood estimator
pdf	probability density function
pmf	probability mass function
GYI	generalized Youden's index
J	Youden's index
ROC	receiver operating characteristic
VUS	volume under the surface
WLOG	without loss of generality

STATISTICAL INFERENCE ON OPTIMAL POINTS TO
EVALUATE MULTI-STATE CLASSIFICATION SYSTEMS

I. Introduction

Decision making occurs daily in a vast range of fields, from health care to information processing and military applications. Generally, these decisions may be based off of classification systems which, for example, label an individual as diseased or not diseased or perhaps label an object of interest as a target or non-target. Although such decisions could be made as simply as through a quick visual inspection, for many decisions of critical importance it is of interest to use statistics and best practices to develop and compare classification systems and quantify their performance so as to choose the best classification method available to aid such decisions [68].

A simple classification rule may classify an item into one of two classes, such as "Positive" and "Negative", or "Diseased" and "Not Diseased". Although a lot of research has been conducted to develop methods for the quantification of such classification systems, most applications in the real world are more complicated and do not fit into simple binary classification rules. Despite examples of classification systems in most applications, this research focuses on examples from a medical diagnostic standpoint, as medical diagnostics carry great importance as well as the possibility for large consequences with respect to misdiagnoses.

One recent example of a medical diagnostic decision involves the use of biomarkers to diagnose subjects post kidney transplant as either being normal kidney function, normal kidney function with proturina (a progression towards the diseased state), or chronic allofraft nephropathy (the diseased state) [58]. Other examples abound such as that of HIV diagnosis. While screening for this disease by using a specific biomarker, patients can be categorized into one of three categories: HIV-negative, HIV-positive non-symptomatic, and HIV-positive with AIDS dementia complex [45]. Extending the health concept to structures, we may be interested in the detection of the stage of structural

damage as being none, within a pre-specified safety range, or beyond the safe operating range. In all of these examples the middle class is important as it represents a state in the progression of some phenomenon (e.g. disease or damage). Thus, diagnosis of the middle class may allow for intervention to prevent a subject or specimen from reaching the end state.

There are methods available to determine the performance of a classification system requiring more than two outcomes. Many of these methods use extensions of receiver operating characteristic (ROC) curve theory for comparing classification systems on their abilities to correctly classify objects [16, 17, 20, 28, 29]. However, the number of possible outcomes is not the only concern when choosing a classification system. The prevalence of the different classes as well as the costs associated with making the correct (or incorrect) decision should also be considered [30, 42, 58, 65]. For example, in HIV diagnosis, different misclassifications may be considered more or less significant. A person who is misdiagnosed as the non-diseased state when they are actually HIV-positive may be considered much worse than the opposite error occurring (a non-diseased person who is diagnosed as HIV-positive). In the first scenario, a person will not receive necessary medical intervention and may now put others at risk since they are unaware of their HIV-positive status. Clearly though, the latter misdiagnosis presents its own cost in that an individual may begin treatment or otherwise suffer with a diagnosis that is incorrect.

In a two-class setting, assigning a cost to the different misclassifications is equivalent to assigning an associated cost to the different correct classifications. However, this equivalence does not universally exist for settings with three or more classes. Currently, little work has been done to compare and quantify the performance of multi-class classification systems by using the misclassifications. By using the misclassifications, different costs may be placed on all the possible errors made by the classification system [58, 65].

The work of this dissertation improves classification system selection and performance quantification for more complicated classification settings involving three or more classes with unequal costs associated with the different misclassification errors. Specifically, precision of estimates of classification system metrics and their optimal points through confidence intervals and hypothesis tests are explored to aid decision makers.

II. Classification and Optimal Performance

2.1 Classification System Families

A classification system (A) is any process that assigns the elements from k partitions of an event set, $\mathbf{E} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_k)$ to k distinct elements of a label set, $\mathbf{L} = (l_1, l_2, ..., l_k)$. These partitions may be referred to as classes. For example, a two-class label set could be {0,1} or {Diseased, Non Diseased}. Data is collected on the elements, which are then processed into a feature or set of features, $\mathbf{F} = (f_1, f_2, ..., f_m)$. These features are then used to assign the different elements from $\mathbf{E}$ to the respective labels, $\mathbf{L}$, $(A : \mathbf{E} \rightarrow \mathbf{F} \rightarrow \mathbf{L})$. It is assumed that there is a parameter or vector of parameters for the features, $\boldsymbol{\theta} \in \Theta$, that can be altered to change the outcome of the classification system.[1] Thus, for every $\boldsymbol{\theta} \in \Theta$, there is a classification system $(A_{\boldsymbol{\theta}})$, and the set of classification systems $\mathbf{A} = (A_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta)$ is called a classification system family (CSF) [58]. It is also assumed that there exists a truth label set, $\mathbf{T} = (t_1, t_2, ..., t_k)$, such that all elements of the population would be correctly labeled by this set.

A two-class classification system has four outcomes with respect to truth (see Table 2.1). Defining one class as positive and the other class as negative, the possible outcomes from the classification system are true positive, true negative, false positive, and false negative. True positive occurs when the system correctly classifies a positive element with a "positive" label (the rate of true positive is often called sensitivity). True negative occurs when the system correctly classifies a negative element with a "negative" label (the rate of true negative is called specificity). These two outcomes are correct classifications. The other two outcomes are misclassifications. A false positive occurs when the system incorrectly classifies a negative element with a "positive" label. Likewise, a false negative occurs when the system incorrectly classifies a positive element with a "negative" label. The results of a classification system are often arranged in a contingency table as seen in Table 2.1 with the truth along the columns and the classification results down the rows.

[1] These parameters will generally be referred to as the thresholds for the classification system.

Table 2.1: Two-class contingency table where green cells correspond to correct classifications and red cells correspond to misclassifications.

		TRUTH	
		Positive	Negative
CLASSIFICATION	"Positive"	True Positive	False Positive
	"Negative"	False Negative	True Negative

An example classification system in a medical diagnostic setting may have elements in partitions of the event set, **E** =(Non-Diseased, Diseased), and the label set, **L** =("Non-Diseased", "Diseased"). After the collection of data such as a patient's blood sample, the feature extracted might be the value of a specific biomarker determined from the blood sample, **F** =(biomarker level, μmol). Then a single threshold, $\theta \in \Theta$, is determined so that whenever the observed biomarker level is less than θ , the patient is labeled as "Diseased", and whenever the biomarker level is greater than θ , the patient is labeled as "Non-Diseased" (see Figure 2.1). For instance, when total cholesterol (a biomarker feature) is greater than 240 (the threshold), a patient may be labeled with "high cholesterol".

In the two-class case, there are two correct classifications and two misclassifications. In the k-class case there are k correct classifications and $k^2 - k$ misclassifications. When there are more than two classes, the correct and misclassifications can no longer be defined as true positive or false negative. Therefore, these terms are generalized to correct classifications and misclassifications. For simplicity of notation, the outcomes are labeled $i \mid j$, where j is the true label for an element and i represents the classification system label for an element, i , $j = 1, 2, ..., k$. Then for all $i = j$, the outcome is a correct classification and for all $i \neq j$, the outcome is a misclassification (see Table 2.2 for $k = 3$).

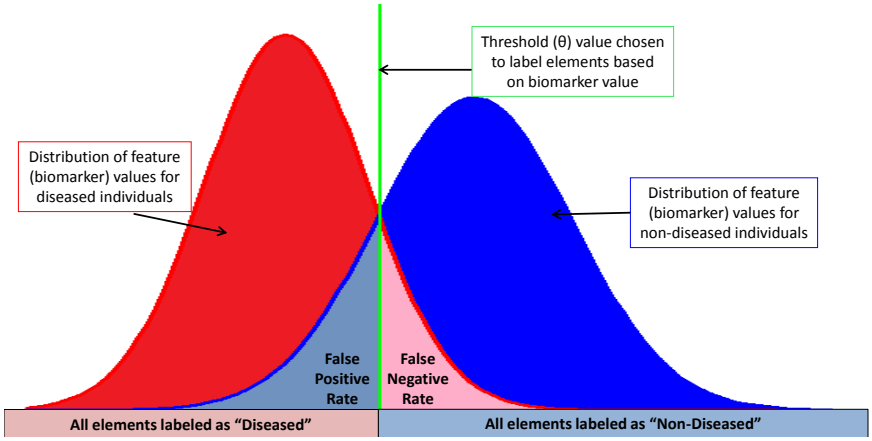

Figure 2.1: Example of a classification system in a medical setting where elements are either diseased or non-diseased. Hypothetical feature distributions for each class and a potential threshold (green line) used to label the elements as either "Diseased" or "Non-Diseased" are shown.

Table 2.2: Three-class contingency table where green cells correspond to correct classifications, $i = j$, and red cells correspond to misclassifications, $i \neq j$.

		TRUTH		
		CLASS 1	CLASS 2	CLASS 3
CLASSIFICATION	"CLASS 1"	1|1	1|2	1|3
	"CLASS 2"	2|1	2|2	2|3
	"CLASS 3"	3|1	3|2	3|3

2.2 Receiver Operating Characteristic Curves

Receiver operating characteristic (ROC) curves are used to describe the performance of a CSF when there are two classes (See Figure 2.2). The ROC curve plots the true positive rate versus the false positive rate over all threshold values, $\theta \in \Theta$. This curve allows for interpretation of the trade-off between the true positive and false positive rates for varying thresholds. Thus, the ROC curve represents the performance of the entire CSF for all $\theta \in \Theta$.

The ROC curve plots classification rates, bounding the curve between 0 and 1 on both the horizontal and vertical axes. The point on the ROC curve that represents perfect classification is (0,1) (Figure 2.2). This point represent a perfect true positive rate (1) and a perfect false positive rate (0). Therefore, CSFs whose ROC curve approach this point are desired and the single classification system closest to this point is optimal. In a two-class setting, the probability of correctly classifying due to random chance is 0.5. The line on the ROC plot that corresponds to chance classification is called the chance line and intersects the points (0,0) and (1,1) (Figure 2.2) [19]. A CSF performing worse than random chance would not be of interest, and therefore only CSFs whose ROC curves lie above the chance line are usually considered. Finally, when there are more than two classes, the ROC curve may be extended to a ROC surface by plotting the correct classification rates over all $\theta \in \Theta$ in a k dimensional space, though only the 3-dimensional surface is visible graphically.

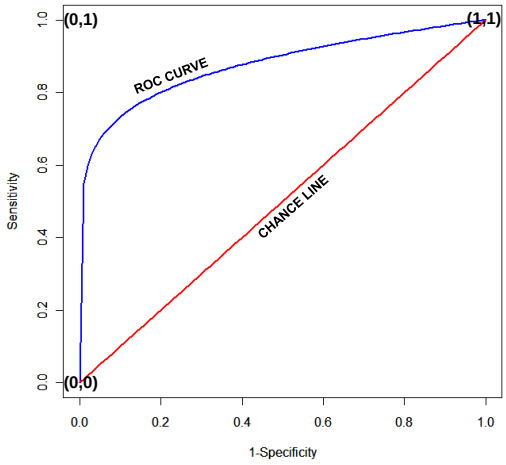

Figure 2.2: Receiver Operating Characteristic Curve.

2.3 Optimal Points

The single classification system resulting in the best classification performance for the CSF is said to occur at the optimal point (or points), corresponding to some $\theta \in \Theta$. For a two-class system, the optimal point is usually found where the probability of a true positive and the probability of a true

negative are maximized (maximization of correct classification probabilities), or equivalently, where the false positive and false negative probabilities are minimized (minimization of misclassification probabilities). Therefore, the optimal point reflects a compromise between the correct classification probabilities (or misclassification probabilities) [42]. The optimal point for a two-class CSF can be found using the ROC curve. If the prevalence of classes and costs associated with classification outcomes are considered equal for both classes, the optimal point occurs where the tangent line to the ROC curve is parallel to the chance line (ie. the slope of the ROC curve is 1) [42]. This is equivalent to finding the point on the ROC curve with the greatest vertical distance from the chance line [54]. The threshold value(s) that produce this point are then chosen as the optimal threshold values for this CSF.

Extensive work in the literature suggests that costs associated with a classification system's outcomes should be taken into account when evaluating the system and estimating optimal thresholds [1, 30, 42, 58, 63–65, 67]. In addition to the costs of the classification outcomes, the prevalence of the different classes may be of importance when determining optimal settings for a CSF [9, 42]. If the *a priori* prevalence of the diseased and non-diseased (or target and non-target) classes as well as the *a priori* costs associated with the decision outcomes are taken into consideration, the CSF may have a different optimal point (see Figure 2.3) [19, 58, 67]. When prevalence and costs are considered, the optimal point occurs on the ROC curve where the slope is equivalent to

$$Slope = \frac{1 - p_P}{p_P} \times \left[\frac{c_{FP} - c_{TN}}{c_{FN} - c_{TP}}\right] \tag{2.1}$$

[42]. The p_P is the prevalence of the positive class, c_{TN} is the cost of a true negative, c_{FP} is the cost of a false positive, c_{TP} is the cost of a true positive, and c_{FN} is the cost of a false negative. Under the assumption of equal prevalences and equal costs of misclassification (or correct classification), this slope is equal to one as expected.

The optimal point for a k-class classification system will usually correspond to at least $k - 1$ threshold values. For example, in order to classify subjects into three categories (HIV negative (NEG), HIV positive non-symptomatic (NAS), and HIV-positive with AIDS dementia complex (ADC)), two threshold values ($\theta_1 < \theta_2$) on a biomarker (NAA/Cr) may be used as a diagnostic

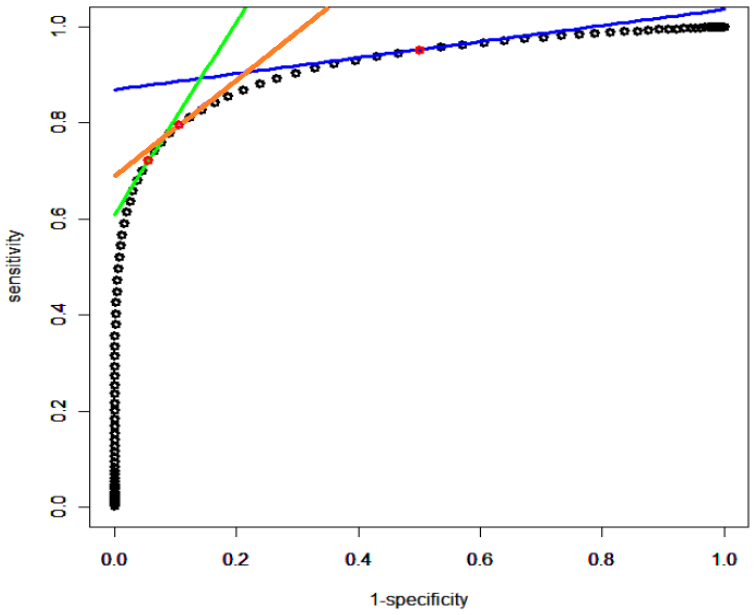

Figure 2.3: Different optimal points (in red) for the same CSF, determined by Equation 2.1. The orange line has a slope of one, representing equal class prevalence and costs associated with the classifications. Both the green and blue lines assume a positive class prevalence of 1/3. The blue line has a slope of 1/6 with $c_{FN} >> c_{FP}$. The green line has a slope of 2 with $c_{FP} = c_{FN}$. For each line $c_{TN} = c_{TP} = 1$.

test [45]. If a subject's NAA/Cr level is below θ_1 they are classified as ADC, if the subject's NAA/Cr level is between θ_1 and θ_2 they are classified as NAS, and finally if the subject's NAA/Cr level is greater than θ_2 they are classified as NEG [45] (see Figure 2.4).

2.4 Metrics for Optimal Points

2.4.1 The Youden Index.

The Youden index (J) was first introduced by W. J Youden in 1950 as an index for rating diagnostic tests (or classification systems) with two classes [76]. The Youden index has been shown to be a useful metric for measuring a classification system's performance as a function of the correct classification probabilities [23, 45, 46, 50, 56, 76]. In a two-class framework, this index is defined

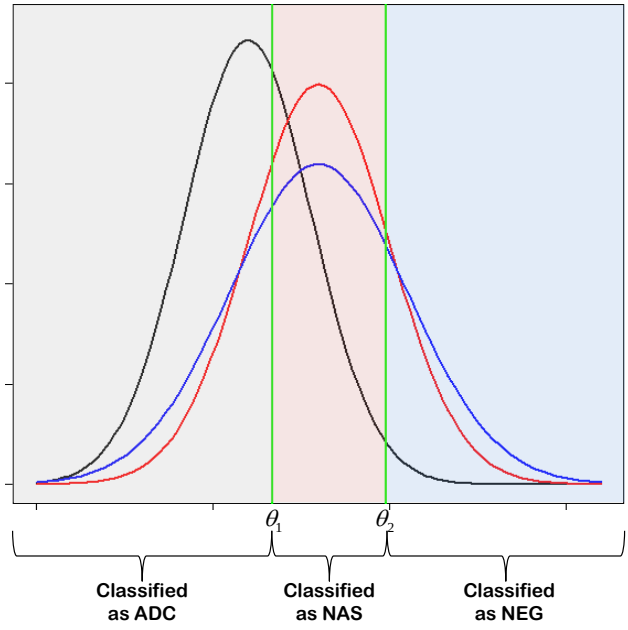

Figure 2.4: Three-class classifications for HIV example. Distributions of the NAA/Cr levels are plotted for ADC (black), NAS (red), and NEG (blue) as well as potential threshold values, θ_1 and θ_2 , used to determine a subject's classification.

as the sum of the system's specificity (true negative rate) and sensitivity (true positive rate) minus one. Using J, the optimal point of the classification system is found by choosing the threshold(s), $\boldsymbol{\theta} \in \Theta$, that maximize J, thereby maximizing the correct classification probabilities. The thresholds associated with the maximum J characterize the CSF at its optimal performance (with respect to correct classification) and correspond to the optimal point on the ROC curve where the slope is equal to one. Therefore, classification systems can be compared by calculating J:

$$J = \max_{\boldsymbol{\theta} \in \Theta} \{sensitivity(\boldsymbol{\theta}) + specificity(\boldsymbol{\theta}) - 1\} \tag{2.2}$$

A classification system which performs worse than chance is generally not of interest, and therefore it is assumed that both sensitivity and specificity are bounded between 0.5 and 1. For this reason, $J = sensitivity(\boldsymbol{\theta}) + specificity(\boldsymbol{\theta}) - 1$ is bounded between 0 and 1 for systems performing better than chance [76].

Costs associated with the different classifications as well as class prevalence may be of importance in the determination of J. In fact, when not explicitly considering a cost structure when using J, a cost and prevalence for each class is being assumed, that of equal weight for all classes [55, 64]. Other costs may be considered by using a generalization to J which incorporates a cost benefit ratio weighted by class prevalence in the two-class framework [30, 63]. The generalized Youden index (GYI) for two classes is defined as

$$GYI = \max_{\boldsymbol{\theta}\in\Theta}\left\{sensitivity(\boldsymbol{\theta}) + \frac{1-p_P}{p_P} \times \left[\frac{c_{FP}-c_{TN}}{c_{FN}-c_{TP}}\right] \times specificity(\boldsymbol{\theta}) - G\right\} \tag{2.3}$$

where G is a constant determined by the prevalence of the positive class and the costs associated with the different decisions [30, 40, 63]. Notice that the prevalence/cost multiplier is the same as in Equation 2.1

When there are more than two classes, J is extended as the sum of the k correct classification probabilities [45, 46]. Under this framework, the correct classification probabilities can no longer be distinguished by sensitivity and specificity, so instead, the k correct classification probabilities are labeled as $P_{i=j|j}(\boldsymbol{\theta})$, where $j = 1, ..., k$ denotes the true class and $i = 1, ..., k$ denotes the classification system's labeled outcome. Then J is redefined as

$$J = \max_{\boldsymbol{\theta}\in\Theta}\left[\sum_{\substack{i=1\\i=j}}^{k}\sum_{j=1}^{k} P_{i|j}(\boldsymbol{\theta})\right] \tag{2.4}$$

J is generalized by adding a multiplier (prevalence and/or utility) to each correct classification probability for classification systems with three or more classes [45, 46]. The limitation with such an extension is that only costs of the total misclassification and (utility of the) total correct classification outcomes within each class are used. This ignores possible different costs on class specific misclassifications. For example, misclassifying stage 3 cancer as stage 2 may have a different cost than classifying stage 3 as stage 1.

Extensive work has derived formulas for determining J and the optimal threshold(s) for CSFs, under various distributional assumptions of the feature used for classification, and focused on the two-class framework [23, 33, 49, 54, 56]. An overview of these results are given in the following sections, and are separated into parametric and nonparametric methods.

2.4.1.1 *Parametric Methods.*

Assume two classes and a single feature used for classification where the feature is independently and normally distributed for each class, where X is the first class and Y is the second class, denote $X_j \sim N(\mu_1, \sigma_1^2)$ for $j = 1, ..., n_1$, $Y_i \sim N(\mu_2, \sigma_2^2)$ for $i = 1, ..., n_2$, and without loss of generality (WLOG) let $\mu_1 < \mu_2$ (see Figure 2.1). Recall that the probability distribution function (pdf) for the normal distribution is

$$f(w \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(w-\mu)^2}{2\sigma^2}} \quad -\infty < w < \infty, \ -\infty < \mu < \infty, \ \sigma > 0 \tag{2.5}$$

Then the Youden index may be written as

$$J = \Phi\left(\frac{\mu_2 - \theta^*}{\sigma_2}\right) + \Phi\left(\frac{\theta^* - \mu_1}{\sigma_1}\right) - 1 \tag{2.6}$$

where Φ is the normal cumulative distribution function (CDF) [56]. Here, the maximum is excluded because the optimal threshold θ^* is used. The closed form solution for the optimal threshold, $\theta^* \in \Theta$, which maximizes Equation 2.6 is given by

$$\theta^* = \frac{\mu_1(b^2 - 1) - a + b\sqrt{a^2 + 2(b^2 - 1)\sigma_1^2 \ln(b)}}{(b^2 - 1)} \tag{2.7}$$

where $a = \mu_2 - \mu_1$ and $b = \frac{\sigma_2}{\sigma_1}$ [56]. If $\sigma_1 = \sigma_2$, this result does not exist, but for this case the optimal point is the midpoint between the distribution means given by [56]:

$$\theta^* = \frac{\mu_1 + \mu_2}{2} \tag{2.8}$$

The GYI may also be rewritten using the normal CDF:

$$GYI = \Phi\left(\frac{\mu_2 - \theta^*}{\sigma_2}\right) + R \times \Phi\left(\frac{\theta^* - \mu_1}{\sigma_1}\right) - G \tag{2.9}$$

where

$$R = \frac{1 - p_p}{p_p} \times \left[\frac{c_{FP} - c_{TN}}{c_{FN} - c_{TP}}\right] \tag{2.10}$$

Again, the maximization is excluded because this equation is being evaluated at the optimal threshold. Accounting for fixed class prevalences and costs associated with the classification outcomes, the optimal threshold when $\sigma_1^2 = \sigma_2^2$ is

$$\theta^* = \frac{2\sigma^2 \ln(R) - \mu_1^2 - \mu_2^2}{2(\mu_2 - \mu_1)} \tag{2.11}$$

[30]. When $\sigma_1^2 \neq \sigma_2^2$ the optimal threshold is

$$\theta^* = \frac{\mu_1(b^2-1) - a + b\sqrt{a^2 + 2(b^2-1)\sigma_1^2 \ln(R \times b)}}{(b^2-1)} \tag{2.12}$$

where $a = \mu_2 - \mu_1$ and $b = \frac{\sigma_2}{\sigma_1}$ [63].

When there are three classes, there is an additional class, Z, where $Z_m \sim N(\mu_3, \sigma_3^2)$ for $m = 1, ..., n_3$. J is then defined as the sum of the three correct classification probabilities and can be expressed using the normal CDF as

$$J = \Phi\left(\frac{\theta_1^* - \mu_1}{\sigma_1}\right) + \Phi\left(\frac{\theta_2^* - \mu_2}{\sigma_2}\right) - \Phi\left(\frac{\theta_1^* - \mu_2}{\sigma_2}\right) - \Phi\left(\frac{\theta_2^* - \mu_3}{\sigma_3}\right) + 1 \tag{2.13}$$

where $\theta_1^* < \theta_2^*$ are the optimal thresholds found to maximize J [45]. The solutions for these optimal thresholds can be found with Equation 2.7 where the solution for θ_1^* is found with $a = \mu_2 - \mu_1$ and $b = \frac{\sigma_2}{\sigma_1}$. The solution for θ_2^* is found similarly with $a = \mu_3 - \mu_2$ and $b = \frac{\sigma_3}{\sigma_2}$ [36]. Although the GYI has not been extended for three classes, in [45] the three-class J is generalized with weights on each correct classification probability. Therefore, weights could be added to Equation 2.13 and the optimal thresholds ($\theta_1^* < \theta_2^*$) would be found numerically.

Finally, for all forms of J and GYI, if the classification feature is distributed log-normally, the point estimate of the threshold is determined using log-transformed data. A similar development is presented in [56] for J with two classes and a gamma distributed feature. However, for features distributed within the Box-Cox family, transformations to normality may be used and the formulas assuming normality applied [30, 45].

2.4.1.2 Nonparametric Methods.

For any number of classes, if no distributional assumptions about the feature used for classification are made, J can be defined using the empirical CDF. The empirical CDF, $F_n(x)$, of a random sample of size n is defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) \tag{2.14}$$

where I is the indicator function and is equal to 1 if the relation is true, and 0 otherwise [32]. For example, in a three-class scenario ($X < Y < Z$), J may be defined as

$$J = \widehat{F}(\theta_1^*) + \widehat{G}(\theta_2^*) - \widehat{G}(\theta_1^*) - \widehat{H}(\theta_2^*) + 1 \tag{2.15}$$

where $\widehat{F}(\theta) = \frac{1}{n_1}\sum_{i=1}^{n_1} I(x_i \leq \theta)$, $\widehat{G}(\theta) = \frac{1}{n_2}\sum_{j=1}^{n_2} I(y_j \leq \theta)$, $\widehat{H}(\theta) = \frac{1}{n_3}\sum_{m=1}^{n_3} I(z_m \leq \theta)$, and θ_1^* and θ_2^* are the thresholds found to maximize Equation 2.15 [45]. Methods that have been used to determine the optimal thresholds include a smoothing kernel method on the empirical CDFs, choosing the observations where the maximum occurs, or by random walks [45, 63].

All forms of J presented may be extended for the k-class J, where again, weights may be placed on the correct classification probabilities to incorporate the importance of the different correct decisions in finding the optimal point [46]. Other work on J includes consideration of special cases such as pooled samples, corrections for measurement error, and methods for when the feature distribution has a mass at zero [49, 54, 55].

2.4.2 Bayes Cost.

The optimal threshold found by maximizing the correct classification probabilities (via J) is equivalent to that found by minimizing the misclassification probabilities in a two-class framework [6, 50]. When there are more than two classes and unequal costs associated with the misclassifications within each class, the equivalence between optimal thresholds found by maximizing correct classification probabilities and minimizing misclassification probabilities is not universally true. This is because it is no longer feasible to assign a simple cost benefit ratio between the benefit of making a correct decision and the costs of making an incorrect decision [58, 65]. Therefore, finding the optimal settings can be more complex when a classification system has more than two classes. In order to assign differing costs or benefits to the potential outcomes of a k-class classification system, a metric that considers all differing misclassification probabilities should be considered instead of extensions of J.

A k-class classification system results in a total of k^2 correct classification and misclassification probabilities; however, J only uses k pieces of information (k correct classification probabilities). Therefore, by using J, $k^2 - k$ pieces of information about the classification system may be lost, namely information about the class-specific misclassifications. A metric developed on the $k^2 - k$ error probabilities will lose no information about the system [58] (see Theorem 1).

For this reason, the development of a metric associated with the misclassification probabilities is of interest. Bayes Cost (BC) is a metric presented in [65] that minimizes misclassification

probabilities for three or more classes. This metric allows for misclassification probabilities to be weighted by the cost and class prevalence associated with each misclassification outcome.

$$Bayes\ Cost = \min_{\boldsymbol{\theta}\in\Theta}\left[\sum_{\substack{i=1\\ i\neq j}}^{k}\sum_{j=1}^{k} c_{i|j}p_j P_{i|j}(\boldsymbol{\theta})\right] \tag{2.16}$$

where $c_{i|j}$ is the fixed cost associated with misclassifying class j as class i and p_j is the fixed prevalence for the j^{th} class. Therefore, *BC* allows for the use of any cost/prevalence structure on both the correct and misclassification probabilities.

Theorem 1. *Using Bayes Cost to determine the optimal thresholds of a multi-state classification system allows for the use of any cost/prevalence structure on any of the correct or misclassification probabilities, therefore not losing any information about the classification system.*

Proof. *Let the prevalence of the class be denoted p_j and the cost of a misclassification be $m_{i\neq j|j}$ or benefit of a correct classification be $b_{i=j|j}$, where the true class is denoted $j = 1, 2, ..., k$ and classification outcomes are denoted $i = 1, 2, ..., k$. The cost function to minimize would be*

$$Cost = \min_{\boldsymbol{\theta}\in\Theta}\left[\sum_{i=1}^{k}\sum_{j=1}^{k} p_j m_{i\neq j|j}P_{i\neq j|j}(\boldsymbol{\theta}) + \sum_{i=1}^{k}\sum_{j=1}^{k} p_j b_{i=j|j}P_{i=j|j}(\boldsymbol{\theta})\right] \tag{2.17}$$

Note, since the classification outcomes in each class are mutually exclusive and the sample size of each class (n_j) is fixed:

$$\sum_{i=1}^{k} P_{i|j}(\boldsymbol{\theta}) = 1, \textit{ for each } j = 1, 2, ..., k \tag{2.18}$$

which implies

$$P_{i=j|j}(\boldsymbol{\theta}) = 1 - \sum_{i=1}^{k} P_{i\neq j|j}(\boldsymbol{\theta}), \textit{ for each } j = 1, 2, ..., k \tag{2.19}$$

Substituting Equation 2.19 in Equation 2.17 gives

$$\begin{aligned}
Cost &= \min_{\boldsymbol{\theta}\in\Theta}\left[\sum_{i=1}^{k}\sum_{j=1}^{k} p_j m_{i\neq j|j} P_{i\neq j|j}(\boldsymbol{\theta}) + \sum_{i=1}^{k}\sum_{j=1}^{k} p_j b_{i=j|j} P_{i=j|j}(\boldsymbol{\theta})\right] \\
&= \min_{\boldsymbol{\theta}\in\Theta}\left[\sum_{i=1}^{k}\sum_{j=1}^{k} p_j m_{i\neq j|j} P_{i\neq j|j}(\boldsymbol{\theta}) + \sum_{i=1}^{k}\sum_{j=1}^{k} p_j b_{i=j|j}\left(\tfrac{1}{k} - P_{i\neq j|j}(\boldsymbol{\theta})\right)\right] \\
&= \min_{\boldsymbol{\theta}\in\Theta}\left[\sum_{i=1}^{k}\sum_{j=1}^{k} p_j m_{i\neq j|j} P_{i\neq j|j}(\boldsymbol{\theta}) + \sum_{i=1}^{k}\sum_{j=1}^{k} \left(\tfrac{p_j}{k} b_{i=j|j} - p_j b_{i=j|j} P_{i\neq j|j}(\boldsymbol{\theta})\right)\right] \\
&= \min_{\boldsymbol{\theta}\in\Theta}\left[\sum_{i=1}^{k}\sum_{j=1}^{k} p_j m_{i\neq j|j} P_{i\neq j|j}(\boldsymbol{\theta}) - \sum_{i=1}^{k}\sum_{j=1}^{k} p_j b_{i=j|j} P_{i\neq j|j}(\boldsymbol{\theta}) + constant\right] \\
&= \min_{\boldsymbol{\theta}\in\Theta}\left[\sum_{i=1}^{k}\sum_{j=1}^{k} p_j m_{i\neq j|j} P_{i\neq j|j}(\boldsymbol{\theta}) - p_j b_{i=j|j} P_{i\neq j|j}(\boldsymbol{\theta})\right] + constant \\
&= \min_{\boldsymbol{\theta}\in\Theta}\left[\sum_{i=1}^{k}\sum_{j=1}^{k} p_j \left(m_{i\neq j|j} P_{i\neq j|j}(\boldsymbol{\theta}) - b_{i=j|j} P_{i\neq j|j}(\boldsymbol{\theta})\right)\right] + constant \\
&= \min_{\boldsymbol{\theta}\in\Theta}\left[\sum_{i=1}^{k}\sum_{j=1}^{k} p_j \left(m_{i\neq j|j} - b_{i=j|j}\right) P_{i\neq j|j}(\boldsymbol{\theta})\right] + constant \\
&= \min_{\boldsymbol{\theta}\in\Theta}\left[\sum_{i=1,i\neq j}^{k}\sum_{j=1}^{k} p_j c_{i|j} P_{i|j}(\boldsymbol{\theta})\right] + constant, \text{ where } c_{i|j} = m_{i\neq j|j} - b_{i=j|j} \\
&= Bayes\ Cost + constant \\
&\implies \boldsymbol{\theta}^*_{Cost} = \boldsymbol{\theta}^*_{BC}
\end{aligned} \tag{2.20}$$

This demonstrates that the optimal thresholds found by minimizing Bayes Cost are equivalent to those found by minimizing a function which uses all classification outcome probabilities from the classification system, allowing for any cost/benefit and prevalence structures to be considered. □

Assume a three-class classification system with a single feature used for classification that is independently and normally distributed for each class, where $\mu_1 < \mu_2 < \mu_3$. Under this framework, BC can be expressed with the standard normal CDF and the optimal thresholds that distinguish between the classes and minimize BC, $\theta_1^* < \theta_2^*$, as:

$$\begin{aligned}
BC_3 = {} & c_{2|1}p_1 \times \left(\Phi\left(\frac{\theta_2^* - \mu_1}{\sigma_1}\right) - \Phi\left(\frac{\theta_1^* - \mu_1}{\sigma_1}\right)\right) + c_{3|1}p_1 \times \left(\Phi\left(\frac{\mu_1 - \theta_2^*}{\sigma_1}\right)\right) \\
& + c_{1|2}p_2 \times \left(\Phi\left(\frac{\theta_1^* - \mu_2}{\sigma_2}\right)\right) + c_{3|2}p_2 \times \left(\Phi\left(\frac{\mu_2 - \theta_2^*}{\sigma_2}\right)\right) \\
& + c_{1|3}p_3 \times \left(\Phi\left(\frac{\theta_1^* - \mu_3}{\sigma_3}\right)\right) + c_{2|3}p_3 \times \left(\Phi\left(\frac{\theta_2^* - \mu_3}{\sigma_3}\right) - \Phi\left(\frac{\theta_1^* - \mu_3}{\sigma_3}\right)\right)
\end{aligned} \tag{2.21}$$

The minimization is not expressed in Equation 2.21 as this is achieved by using the optimal thresholds. The optimal thresholds must be found numerically when all $c_{i|j}p_j$ are not equal, for $i \neq j$. When all $c_{i|j}p_j$ are equal, for $i \neq j$, Equation 2.7 may be used to find the optimal threshold

between each set of normal distributions. Equation 2.21 may be extended for any k classes with a single feature used for classification that is independently and normally distributed for each class, and would require $k-1$ optimal thresholds.

When there are two classes, the optimal threshold found by minimizing BC is equivalent to that found by maximizing the GYI, Equations 2.11 or 2.12 (assuming the same costs and prevalences used to find the optimal point). A proof of this equivalence is given in Section 2.5.2. Also, if all $c_{i|j}p_j$ are equal, for $i \neq j$, the optimal threshold(s) found with BC would be equivalent to those found by maximizing J.

In a nonparametric setting, BC can be estimated using the empirical distribution function. Letting $\boldsymbol{\theta} = (\theta_1 < \theta_2 < ... < \theta_k)$ and F_j be the empirical CDF for the j^{th} class with $F_{j-1} < F_j$, for all k classes, BC is defined as

$$BC = \min_{\theta \in \Theta} \left[\sum_{\substack{i=1 \\ i \neq j}}^{k} \sum_{j=1}^{k} c_{i|j} p_j \left[F_j(\theta_i) - F_j(\theta_{i-1}) \right] \right] \tag{2.22}$$

where $F_j(\theta_0) = 0$ and $F_j(\theta_k) = 1$ [65]. The optimal thresholds are then found to be those which minimize Equation 2.22.

2.5 Confidence on Optimal Point Metrics

It is critical to characterize the uncertainty in an optimal point, as such estimates are typically constructed from data. This is most commonly accomplished by creating confidence intervals (CIs) around the metric used to characterize the optimal point (Youden index, Bayes Cost, etc) as well as creating confidence interval(s) around the threshold(s) which correspond to the optimal point [30, 33, 49, 56, 76].

CIs are a statistical inference method that provide a range of values (usually an interval) for which there is a specified level of confidence that the true parameter lies within the interval. CIs may be constructed as either one or two sided (one sided being of the form where there is either a lower or upper bound, but not both). This work focuses on constructing two sided confidence intervals. If $\mathbf{X} = (X_1, \ldots, X_n)$ is a random sample, then $L(\mathbf{X})$ and $U(\mathbf{X})$ form a confidence interval with confidence coefficient $1-\alpha$ for some function of the parameter θ , $\tau(\theta)$, such that $P[L(\mathbf{X}) \leq \tau(\theta) \leq U(\mathbf{X})] = 1-\alpha$ [12, p. 417],[44, p.377]. Because it is known that the upper

and lower bounds of the CI are functions of the observed data, the notation for the bounds may be simplified by writing $L(\mathbf{X})$ as $\tau(\theta)_L$ and $U(\mathbf{X})$ as $\tau(\theta)_U$.

Not all CIs perform equally well. An interval's coverage probability and length are metrics of a CI's performance. If a CI with a confidence coefficient of $1-\alpha$ is constructed 100 times, it is expected that $(1-\alpha)100\%$ of the intervals actually contain the true parameter of interest. This may not always be the case, and the percent of constructed CIs that contain the true parameter is the coverage probability of the CI. The coverage probability should be at least $(1-\alpha)100\%$ for a well performing CI. CIs with coverage probability greater than $(1-\alpha)100\%$ are considered conservative.

For all CIs that meet the desired coverage probability, it is then of interest to find the interval with the shortest length. The length of an interval is defined as $\tau(\theta)_U - \tau(\theta)_L$. A shorter length CI which meets the desired coverage probability provides a more precise (and therefore, arguably, a more useful) estimate of the parameter. Another metric of CI performance is its symmetry, which may be used to judge whether or not the true parameter of interest lies in the center of the interval, or if the interval is skewed to one side. Mean squared error and bias of the parameter estimate may impact CI performance and are therefore also sometimes considered, though these are not properties of the interval itself.

2.5.1 Confidence on the Youden Index and Optimal Thresholds.

Several methods exist in the literature for constructing CIs around J and the optimal threshold(s), mainly in a two-class setting. In addition to these methods, bootstrap methods are also applicable, as bootstrap CIs are a general and flexible method that may be used under any distributional assumptions of the features and classification system structure. First, parametric CI methods are presented and following these methods, the nonparametric CI methods available for J are presented.

A delta method approximation, which uses first order Taylor series expansions to determine the variance of J and the optimal threshold(s), has been implemented to create CIs for J and the resulting optimal threshold(s) for a classification system with two or three classes with a single feature that is independently and normally distributed for each class [36, 56, 63, 64]. The delta

method $(1-\alpha)100\%$ CI around J is

$$\widehat{J} \pm z_{\alpha/2}\sqrt{Var\left(\widehat{J}\right)} \tag{2.23}$$

where $\widehat{J}$ is estimated using Equation 2.6 for two classes or Equation 2.13 for three classes and $Var\left(\widehat{J}\right)$ is approximated with the delta method as:

$$Var\left(\widehat{J}\right) \approx \sum_{j=1}^{k}\left[\left(\frac{\partial J}{\partial \mu_j}\right)^2 Var(\widehat{\mu_j}) + \left(\frac{\partial J}{\partial \sigma_j}\right)^2 Var(\widehat{\sigma_j})\right] \tag{2.24}$$

The covariance term from the delta method approximation is zero, due to the assumption of independence between the classes' feature distributions.

Assuming two classes and a normally distributed feature, $X_j \sim N(\mu_1, \sigma_1^2)$ for $j = 1, ..., n_1$, $Y_i \sim N(\mu_2, \sigma_2^2)$ for $i = 1, ..., n_2$, and $\mu_1 < \mu_2$, $Var\left(\widehat{J}\right)$ in Equation 2.24 is estimated by:

$$\begin{aligned} Var\left(\widehat{J}\right) \approx & \left(\frac{S_2^2}{n_2}\right)\left[\phi(\widehat{z_2}) + (\phi(\widehat{z_1}) - \phi(\widehat{z_2}))\left(\frac{-1+\widehat{a}\,\widehat{b}(rad)^{-1/2}}{\widehat{b}^2-1}\right)\right]^2 \\ & + (-1)\left(\frac{S_1^2}{n_1}\right)\left[\phi(\widehat{z_1}) + (\phi(\widehat{z_1}) - \phi(\widehat{z_2}))\left(\frac{\widehat{b}^2+(-1)\widehat{a}\,\widehat{b}(rad)^{-1/2}}{\widehat{b}^2-1}\right)\right]^2 \\ & + \frac{1}{2(n_2-1)}\begin{bmatrix} \widehat{z_2}\phi(\widehat{z_2}) + \frac{(\widehat{b}\phi(\widehat{z_1})-\phi(\widehat{z_2}))}{(\widehat{b}^2-1)^2(S_1^2)^{1/2}} \\ \times(2\widehat{a}\,\widehat{b}^2 + ((-\widehat{b}^2-1)(rad)^{1/2} \\ +(S_2^2)(\widehat{b}^2-1)(rad)^{-1/2}(\ln(\widehat{b}^2) + 1 - \widehat{b}^{-2}))) \end{bmatrix}^2 \\ & + \frac{-1}{2(n_1-1)}\begin{bmatrix} \widehat{z_1}\phi(\widehat{z_1}) + \frac{(\widehat{b}\phi(\widehat{z_1})-\phi(\widehat{z_2}))}{(\widehat{b}^2-1)^2(S_1^2)^{1/2}} \\ \times(2\widehat{a}\,\widehat{b}^2 + ((-\widehat{b}^2-1)(rad)^{1/2} \\ +(S_1^2)(\widehat{b}^2-1)(rad)^{-1/2}(\ln(\widehat{b}^2) + \widehat{b}^2 - 1))) \end{bmatrix}^2 \end{aligned} \tag{2.25}$$

where $\widehat{z_2} = \frac{\overline{x}-\widehat{\theta^*}}{\sqrt{S_2^2}}$, $\widehat{z_1} = \frac{\widehat{\theta^*}-\overline{y}}{\sqrt{S_1^2}}$, $\widehat{a} = \overline{y} - \overline{x}$, $\widehat{b} = \frac{S_2^2}{S_1^2}$, $rad = \widehat{a}^2 + \left(\widehat{b}^2 - 1\right)S_1^2\ln\left(\widehat{b}^2\right)$, and ϕ represents the standard normal pdf [56]. A similar formulation of the approximation of $Var\left(\widehat{J}\right)$ is used for the delta method CI for the three-class J.

The $(1-\alpha)100\%$ CI(s) for the optimal threshold(s) (two or three classes with a normally distributed feature) is given by

$$\theta^* \pm z_{\alpha/2}\sqrt{Var(\widehat{\theta^*})} \tag{2.26}$$

where $\widehat{\theta^*}$ may be found with Equation 2.7 (for either optimal threshold by considering the appropriate adjacent classes) [36, 56]. Using the delta method, the variance of $\widehat{\theta^*}$ is approximated with

$$Var(\widehat{\theta^*}) \approx \left(\frac{\partial\theta^*}{\partial\mu_1}\right)^2 Var(\widehat{\mu_1}) + \left(\frac{\partial\theta^*}{\partial\sigma_1}\right)^2 Var(\widehat{\sigma_1}) + \left(\frac{\partial\theta^*}{\partial\mu_2}\right)^2 Var(\widehat{\mu_2}) + \left(\frac{\partial\theta^*}{\partial\sigma_2}\right)^2 Var(\widehat{\sigma_2}) \tag{2.27}$$

The partial derivatives required for this approximation are

$$\left(\frac{\partial\theta^*}{\partial\mu_1}\right) = \frac{b^2 + ab(rad)^{-1/2}(-1)}{b^2 - 1} \tag{2.28}$$

$$\left(\frac{\partial\theta^*}{\partial\mu_2}\right) = \frac{b^2 + ab(rad)^{-1/2}(-1)}{b^2 - 1} \tag{2.29}$$

$$\left(\frac{\partial\theta^*}{\partial\sigma_1}\right) = \frac{-2ab^2}{(b^2-1)^2\sigma_1} + \left[\frac{b(b^2+1)(rad)^{1/2}}{(b^2-1)^2\sigma_1} - \frac{\sigma_1 b(rad)^{-1/2}}{b^2-1}\left(\ln(b^2) + b^2 - 1\right)\right] \tag{2.30}$$

$$\left(\frac{\partial\theta^*}{\partial\sigma_2}\right) = \frac{-2ab^2}{(b^2-1)^2\sigma_1} + \left[\frac{(-b^2-1)(rad)^{1/2}}{(b^2-1)^2\sigma_1} + \frac{\sigma_2 b(rad)^{-1/2}}{b^2-1}\left(\ln(b^2) + 1 - b^{-2}\right)\right] \tag{2.31}$$

where a, b, and rad are defined as they were for Equation 2.25 [36, 56]. When there are three classes, the variance and partial derivatives of the second optimal threshold are estimated with Equation 2.27 and Equations 2.28 to 2.31 by replacing the first class with the second and the second class with third [36, 56].

Under the framework of the two-class GYI, the delta method has been used for developing a CI around the optimal threshold for a classification system which utilizes a single normally or log-normally distributed feature (but not for the GYI itself) [30]. For a CI around the optimal threshold found with the GYI, the delta method CI is similarly developed as that for J, although the expression allows for the cost/benefit weighting factor. When the variances are equal

$$Var(\widehat{\theta^*}) \approx \left(\frac{\ln(R)}{a}\right)^2 Var(\widehat{\sigma^2}) + \left(1/2 + \frac{\sigma^2\ln(R)}{a^2}\right)^2 Var(\widehat{\mu_1}) + \left(1/2 - \frac{\sigma^2\ln(R)}{a^2}\right)^2 Var(\widehat{\mu_2}) \tag{2.32}$$

This approximation may be used in Equation 2.26 to construct the CI around the optimal threshold [30]. This CI has also been generalized for when the variances are not equal [64]. Further, the delta method has been used to derive CIs for J and the optimal threshold when the classification system utilizes a single feature to distinguish between two classes when the distribution of the feature for each class is an independent gamma [56].

In [43], the delta method CIs for the two-class J and the optimal threshold are modified by utilizing a second order Taylor series expansion as opposed to the first order expansion used in Equations 2.24 and 2.27. Although the extension to the delta method is presented, the performance of the extended version is not compared to the simpler method and therefore the more complicated derivation has not been justified. All delta method CIs are only appropriate for large sample sizes if the desired coverage probability is to be achieved.

Generalized CIs (GCIs) are developed in [33] for J and the optimal threshold under the assumption of a single feature used for the classification between two classes, where the feature is independently and normally distributed for each class. These exact CIs outperform the delta method CIs for scenarios considered in the simulation presented in [33] because they meet the desired coverage (for small $n_j \geq 10$) while maintaining a CI length that is less than the delta method CI length. This generalized method for classes with a normally distributed feature is also used for constructing a CI on the difference in paired Youden indices in the two-class framework, allowing for the comparison of two classification systems' performances in a paired data structure [80].

If no assumptions are made about the distribution of the feature used for classification, a nonparametric CI around J and the optimal threshold may be used. In [79], a CI for the two-class J is developed with the Agrestti-Coull confidence interval for a binomial proportion (see [2]), where J is estimated with

$$\widehat{J} = \frac{\sum_{i=1}^{n_1} I(X_i \leq \theta^*) + z_{1-\alpha/2}^2/2}{n_1 + z_{1-\alpha/2}^2/2} - \frac{\sum_{j=1}^{n_2} I(Y_j \leq \theta^*) + z_{1-\alpha/2}^2/2}{n_2 + z_{1-\alpha/2}^2/2} \tag{2.33}$$

A nonparametric asymptotic normal (AN) bootstrap is utilized to determine the CI bounds for J estimated in Equation 2.33 (this method does not provide a CI for the optimal threshold). Under various distributional assumptions, this method approaches the desired coverage probability for $n_j \geq 50$. In [43], an empirical likelihood method which utilizes bootstraps is used for constructing a nonparametric CI around J and the optimal threshold in the two-class framework. This nonparametric method performs well with respect to coverage for samples of at least 30 in each class.

Currently, a confidence interval around the GYI has not been presented, except for a bootstrap CI [40].

2.5.2 *Confidence on Bayes Cost and Optimal Thresholds.*

The two-class J may be written as

$$J = \max_{\boldsymbol{\theta}\in\Theta} \left[P_{1|1}(\boldsymbol{\theta}) + P_{2|2}(\boldsymbol{\theta}) - 1 \right] \tag{2.34}$$

where $P_{1|1}(\boldsymbol{\theta})$ and $P_{2|2}(\boldsymbol{\theta})$ are the correct classification probabilities for a threshold(s), $\boldsymbol{\theta} \in \Theta$. The two-class BC (with all $c_{i|j}p_j$ assumed to be one, for $i \neq j$) may be written

$$BC = \min_{\boldsymbol{\theta}\in\Theta} \left[P_{2|1}(\boldsymbol{\theta}) + P_{1|2}(\boldsymbol{\theta}) \right] \tag{2.35}$$

where $P_{2|1}(\boldsymbol{\theta})$ and $P_{1|2}(\boldsymbol{\theta})$ are the misclassification probabilities for a $\boldsymbol{\theta} \in \Theta$. For greater utility, BC is defined with prevalences on the two classes and different costs on misclassification errors [58, 65]:

$$BC = \min_{\boldsymbol{\theta}\in\Theta} \left[c_{2|1}p_1 P_{2|1}(\boldsymbol{\theta}) + c_{1|2}p_2 P_{1|2}(\boldsymbol{\theta}) \right] \tag{2.36}$$

where $c_{i|j}$ is the fixed cost associated with misclassifying class j as class i and p_j is the fixed prevalence for the j^{th} class.

From these definitions it is shown that for a two-class classification system, the optimal threshold found by minimizing BC is equivalent to the optimal threshold found by maximizing J (when all $c_{i|j}p_j$ are equal, for $i \neq j$, Theorem 2) or by maximizing the GYI (Theorem 3) when the costs are defined as

$$\frac{c^{GYI}_{1|2} - c^{GYI}_{2|2}}{c^{GYI}_{2|1} - c^{GYI}_{1|1}} = \frac{c^{BC}_{1|2}}{c^{BC}_{2|1}} \tag{2.37}$$

where $c^{GYI}_{i|j}$ and $c^{BC}_{i|j}$ are the costs associated with the GYI and BC, respectively. Then, a CI around the optimal threshold found by minimizing BC would be equivalent to the CIs developed for the optimal threshold found with J or the GYI (assuming the same statistical method for constructing the CI is used).

Theorem 2. *The optimal threshold, $\boldsymbol{\theta}^*_{BC}$, found by minimizing Bayes Cost when all $c_{i|j}p_j$ are assumed equal to one, for $i \neq j$, is equivalent to the optimal threshold found by maximizing the Youden index, $\boldsymbol{\theta}^*_J = \boldsymbol{\theta}^*_{BC}$, for a two-class classification system family.*

Proof. *Let $\boldsymbol{\theta}^*_{BC}$ and $\boldsymbol{\theta}^*_J$ represent the optimal thresholds found by minimizing Bayes Cost and maximizing the Youden index, respectively. Also, let $P_{i|j}(\boldsymbol{\theta})$ represent the probability of classifying*

class j as class i . Then, there exits $\boldsymbol{\theta}^*_{BC} \ni BC = \min_{\boldsymbol{\theta}\in\Theta}\left[P_{2|1}(\boldsymbol{\theta}) + P_{1|2}(\boldsymbol{\theta})\right]$ and there exists $\boldsymbol{\theta}^*_J \ni J = \max_{\boldsymbol{\theta}\in\Theta}\left[P_{1|1}(\boldsymbol{\theta}) + P_{2|2}(\boldsymbol{\theta}) - 1\right]$. Now, consider

$$\begin{aligned}
\boldsymbol{\theta}^*_{BC} &= \arg\min_{\boldsymbol{\theta}\in\Theta}\left[P_{2|1}(\boldsymbol{\theta}) + P_{1|2}(\boldsymbol{\theta})\right] \\
&= \arg\max_{\boldsymbol{\theta}\in\Theta}\left[1 - P_{2|1}(\boldsymbol{\theta}) - P_{1|2}(\boldsymbol{\theta})\right] \\
&= \arg\max_{\boldsymbol{\theta}\in\Theta}\left[1 - (1 - P_{1|1}(\boldsymbol{\theta})) - (1 - P_{2|2}(\boldsymbol{\theta}))\right] \\
&= \arg\max_{\boldsymbol{\theta}\in\Theta}\left[1 - 1 + P_{1|1}(\boldsymbol{\theta}) - 1 + P_{2|2}(\boldsymbol{\theta})\right] \\
&= \arg\max_{\boldsymbol{\theta}\in\Theta}\left[P_{1|1}(\boldsymbol{\theta}) + P_{2|2}(\boldsymbol{\theta}) - 1\right] \\
&= \boldsymbol{\theta}^*_J \\
\Rightarrow\ & \boldsymbol{\theta}^*_{BC} = \boldsymbol{\theta}^*_J
\end{aligned} \tag{2.38}$$

□

Theorem 3. *The optimal threshold, $\boldsymbol{\theta}^*_{BC}$, found by minimizing Bayes Cost is equivalent to the optimal threshold found by maximizing the generalized Youden index, $\boldsymbol{\theta}^*_{GYI} = \boldsymbol{\theta}^*_{BC}$, for a two-class classification system when the costs are defined where $\left[(c^{GYI}_{1|2} - c^{GYI}_{2|2})/(c^{GYI}_{2|1} - c^{GYI}_{1|1})\right] = \left[c^{BC}_{1|2}/c^{BC}_{2|1}\right]$.*

Proof. *Let $\boldsymbol{\theta}^*_{BC}$ and $\boldsymbol{\theta}^*_{GYI}$ represent the optimal thresholds found by minimizing BC and maximizing the GYI, respectively. Also, let $c_{i|j}$ be the fixed cost associated with classifying class j as class i , p_j be the fixed prevalence for the j^{th} class, and $P_{i|j}(\boldsymbol{\theta})$ be the probability of classifying class j as class i for a given $\boldsymbol{\theta} \in \Theta$. Assume $\left[(c^{GYI}_{1|2} - c^{GYI}_{2|2})/(c^{GYI}_{2|1} - c^{GYI}_{1|1})\right] = \left[c^{BC}_{1|2}/c^{BC}_{2|1}\right]$, then there exits $\boldsymbol{\theta}^*_{GYI} \ni GYI = \max_{\boldsymbol{\theta}\in\Theta}\left[sensitivity(\boldsymbol{\theta}) + \frac{1-p_p}{p_p} \times \left[\frac{c^{GYI}_{2|2} - c^{GYI}_{1|2}}{c^{GYI}_{1|1} - c^{GYI}_{2|1}}\right] \times specificity(\boldsymbol{\theta}) - 1\right]$ and there exists*

$\boldsymbol{\theta}^*_{BC} \ni BC = \min_{\boldsymbol{\theta}\in\Theta}\left[p_1 c^{BC}_{2|1} P_{2|1}(\boldsymbol{\theta}) + p_2 c^{BC}_{1|2} P_{1|2}(\boldsymbol{\theta})\right]$. *Then*

$$
\begin{aligned}
\boldsymbol{\theta}^*_{GYI} &= \arg\max_{\boldsymbol{\theta}\in\Theta}\left[sensitivity(\boldsymbol{\theta}) + \tfrac{1-p_p}{p_p} \times \left[\tfrac{c^{GYI}_{2|2} - c^{GYI}_{1|2}}{c^{GYI}_{1|1} - c^{GYI}_{2|1}}\right] \times specificity(\boldsymbol{\theta}) - 1\right] \\
&= \arg\max_{\boldsymbol{\theta}\in\Theta}[P_{1|1}(\boldsymbol{\theta}) + \tfrac{1-p_1}{p_1} \times \left[\tfrac{c^{GYI}_{2|2} - c^{GYI}_{1|2}}{c^{GYI}_{1|1} - c^{GYI}_{2|1}}\right] \times P_{2|2}(\boldsymbol{\theta}) - G] \\
&= \arg\max_{\boldsymbol{\theta}\in\Theta}\left[P_{1|1}(\boldsymbol{\theta}) + \tfrac{p_2}{p_1} \times \left[\tfrac{c^{GYI}_{2|2} - c^{GYI}_{1|2}}{c^{GYI}_{1|1} - c^{GYI}_{2|1}}\right] \times P_{2|2}(\boldsymbol{\theta}) - G\right] \\
&= \arg\max_{\boldsymbol{\theta}\in\Theta}\left[(1 - P_{2|1}(\boldsymbol{\theta})) + \tfrac{p_2}{p_1} \times \left[\tfrac{c^{GYI}_{2|2} - c^{GYI}_{1|2}}{c^{GYI}_{1|1} - c^{GYI}_{2|1}}\right] \times (1 - P_{1|2}(\boldsymbol{\theta}))\right] \\
&= \arg\max_{\boldsymbol{\theta}\in\Theta}\left[-P_{2|1}(\boldsymbol{\theta}) + \tfrac{p_2}{p_1} \times \left[\tfrac{c^{BC}_{1|2}}{c^{BC}_{2|1}}\right] \times (1 - P_{1|2}(\boldsymbol{\theta}))\right] \\
&= \arg\max_{\boldsymbol{\theta}\in\Theta}\left[-P_{2|1}(\boldsymbol{\theta}) + \tfrac{p_2}{p_1} \times \left[\tfrac{c^{BC}_{1|2}}{c^{BC}_{2|1}}\right] - \tfrac{p_2}{p_1} \times \left[\tfrac{c^{BC}_{1|2}}{c^{BC}_{2|1}}\right] \times P_{1|2}(\boldsymbol{\theta})\right] \\
&= \arg\max_{\boldsymbol{\theta}\in\Theta}\left[\tfrac{1}{p_1 c^{BC}_{2|1}}\left(-p_1 c^{BC}_{2|1} P_{2|1}(\boldsymbol{\theta}) - p_2 c^{BC}_{1|2} P_{1|2}(\boldsymbol{\theta}) + p_2 c^{BC}_{1|2}\right)\right] \\
&= \arg\max_{\boldsymbol{\theta}\in\Theta}\left[\tfrac{1}{constant}\left(-p_1 c^{BC}_{2|1} P_{2|1}(\boldsymbol{\theta}) - p_2 c^{BC}_{1|2} P_{1|2}(\boldsymbol{\theta}) + constant\right)\right] \\
&= \arg\max_{\boldsymbol{\theta}\in\Theta}\left[-p_1 c^{BC}_{2|1} P_{2|1}(\boldsymbol{\theta}) - p_2 c^{BC}_{1|2} P_{1|2}(\boldsymbol{\theta})\right] \\
&= \arg\min_{\boldsymbol{\theta}\in\Theta}\left[p_1 c^{BC}_{2|1} P_{2|1}(\boldsymbol{\theta}) + p_2 c^{BC}_{1|2} P_{1|2}(\boldsymbol{\theta})\right] \\
&= \boldsymbol{\theta}^*_{BC} \\
\Rightarrow\ & \boldsymbol{\theta}^*_{GYI} = \boldsymbol{\theta}^*_{BC}
\end{aligned}
\tag{2.39}
$$

□

A delta method CI for the optimal threshold found by minimizing *BC* is presented in [63] for a classification system with two classes and a single feature that is independently and normally distributed for each class. This CI is equivalent to the delta method CI for the optimal threshold found with the GYI (Section 2.5.1) when costs are defined with Equation 2.37.

CIs on the optimal thresholds found by minimizing *BC* in a multi-state setting are derived using the delta method and numerical approximations in [65]. Notably, the CI for *BC* was not derived. However, in a three-class scenario, CIs on the two threshold values may not necessarily correspond to confidence around the optimal point. A specific set of thresholds $\{\theta_1,\theta_2\}$ from the CIs around each individual threshold may be a hidden extrapolation outside the optimal threshold region. Therefore, CIs around the optimal thresholds may not be the ideal method for quantifying uncertainty in the optimal point, especially in a multi-state setting with more than one threshold. To quantify uncertainty in the optimal point, CIs around the optimal point metric (*J* or *BC*) should be considered.

To further motivate a CI around the optimal point metric as opposed to the optimal thresholds only, consider the following example. Assume a random draw from a classification system with three classes (assume samples of size 50 are taken from each class), where $X_1 \sim N(-3, 1)$, $X_2 \sim N(0, 1)$, and $X_3 \sim N(3, 1)$. The delta method CIs around the two optimal thresholds found to distinguish between the classes may be $\theta_1 \in [-1.95, -1.49]$ and $\theta_2 \in [1.45, 2.12]$ using the method in [65]. Given the estimated normal distributions from the sample, this range of thresholds would correspond to BC values from 0.207 to 0.226. However, for the same sample, the delta method CI around BC (developed in Section 3.2) is $BC \in [0.114, 0.301]$ and the true value of BC from the assumed underlying distributions is 0.27. Therefore, values within the thresholds' CIs do not necessarily reflect all the uncertainty in the optimal performance of the system (measured by BC), and in this example, overestimates the system's performance.

Notably, the CIs around the thresholds are of use once a classification system has been chosen for implementation. Before a classification system is chosen, however, it may be of interest to compare competing systems based on their optimal performance in order to chose the system with the most powerful classification ability. By constructing a CI around each classification system's BC value, performance at the optimal settings can be compared between systems. Currently, methods for CIs around BC do not exist.

2.6 Hypothesis Tests for Optimal Point Metrics

A hypothesis "is a statement about a population parameter" [12, p. 373]. In testing a hypothesis there are two hypotheses, the null hypothesis ($H_0, \theta \in \Theta_0$) and the alternate hypothesis ($H_1, \theta \in \Theta_0^C$). Both of these hypotheses make statements about the parameter space of interest, such that combined, they cover the entire parameter space [34, p.60]. Of interest for this work would be hypotheses about metrics of a classification system, such as J or BC. Such a hypothesis might be constructed to test if a classification system meets some desired level of performance, for instance, to determine if a classification system performs better than chance.

There are two types of errors which may occur when testing a hypothesis. A Type I error occurs if the null hypothesis is rejected, when it was actually true (ie. $\theta \in \Theta_0$). A Type II error occurs when the null hypothesis is not rejected, when it is not true (ie. $\theta \in \Theta_0^C$). Clearly, it is ideal to minimize the

probability of committing either of the two errors. However, there is a trade-off between both errors. Therefore, a level of significance of the test ($\alpha \in [0, 1]$) is usually set such that the probability of a Type I error is less than or equal to the level of significance for all $\theta \in \Theta_0$ [34, p.61]. Then for all tests with the desired level of significance, the test which minimizes the probability of a Type II error would be best.

Although tests of hypotheses on J or BC would be useful when selecting a classification system, such tests have not been developed.

2.7 Distributions for the Youden Index and Bayes Cost Inference

Making no distributional assumptions about a classification system, the classification outcomes with respect to truth can be modeled as binomial or multinomial random variables, for $k = 2$ or $k \geq 3$ classes, respectively. Therefore, background information on these distributions is presented in this section.

2.7.1 Binomial Distribution.

The classification outcomes from a two-class classification system, for a fixed $\boldsymbol{\theta} \in \Theta$, are arranged in a contingency table in Table 2.3, where $X_{i|j}$ denotes the number of observations classified into class i with truth class j. The sample drawn from each class is fixed; consequently, the knowledge of the total correct or incorrect observations explicitly defines the other. For that reason, the correct or incorrect classification observations from each class are modeled as binomial$(n_j, p_{i|j})$, where $p_{i|j}$ is the true population probability for the outcome of interest and n_j is the fixed number sampled from the j^{th} class, $j = 1, 2$. The binomial probability mass function (pmf) is given, generally, by

$$f_X(x \mid n, p) = P(X = x|n, p) = \binom{n}{x} p^x (1-p)^{n-x} \quad x = 0, 1, ...n, \;\; 0 \leq p \leq 1 \tag{2.40}$$

The maximum likelihood estimate (MLE) of p, is $\widehat{p} = \frac{x}{n}$.

2.7.1.1 Confidence Interval for Binomial Proportions.

Clopper and Pearson derived an exact CI for a binomial probability using fiducial limits in 1934 [13]. The Clopper-Pearson $(1 - \alpha)100\%$ CI for p from an observed statistic, y = number of successes, from a binomial distribution with $f_Y(y \mid p)$ defined as the binomial pmf, is found by

Table 2.3: Contingency table for a two-class classification system. Column labels represent truth and row labels represent the label given by the classification system.

	Class 1	Class 2
Test = 1	$X_{1\|1}$	$X_{1\|2}$
Test = 2	$X_{2\|1}$	$X_{2\|2}$

solving the following two equations for the lower and upper bound (p_L and p_U, respectively)

$$\sum_{k=y}^{n} f_Y(k \mid p_L) = \sum_{k \geq y} f_Y(k \mid p_L) = \frac{\alpha}{2} \tag{2.41}$$

$$\sum_{k=0}^{y} f_Y(k \mid p_U) = \sum_{k \leq y} f_Y(k \mid p_U) = \frac{\alpha}{2} \tag{2.42}$$

The sample space is $y \in (0, ..., n)$. When $y = 0$ or $y = n$ is observed, a solution cannot be found for one of the two above equations (2.41 and 2.42) and the lower bound is 0 or the upper bound is 1, respectively [3, p.18]. This last condition is necessary because these extreme values of Y result in either summation for any p to be 1, due to the property of a pmf where

$$\sum_{y \in \mathcal{Y}} f_Y(y \mid \theta) = 1 \tag{2.43}$$

for any θ. The closed form solution of the Clopper-Pearson interval for a binomial probability is

$$\left[1 + \frac{n - x + 1}{x F_{2x,2(n-x+1),1-\alpha/2}}\right]^{-1} < p < \left[1 + \frac{n - x}{(x+1) F_{2(x+1),2(n-x),\alpha/2}}\right]^{-1} \tag{2.44}$$

where x is the observed number of successes ($x = 1, 2, \ldots, n-1$) and this interval has a coverage probability of at least $(1 - \alpha)100\%$ for all p [2].

2.7.2 Multinomial Distribution.

A $k \times k$ contingency table is used for arranging the outcomes of a k-class classification system for a fixed $\boldsymbol{\theta} \in \Theta$ (Table 2.4). The multivariate random variable $\mathbf{X_j} = (X_{1|j}, X_{2|j}, ..., X_{k|j})$ represents the k outcomes from a single class sampled n_j times and is distributed multinomial($n_j, \mathbf{p_j} = (p_{1|j}, p_{2|j}, ..., p_{k|j})$) where $p_{i|j}$ represents the true probability for the j^{th} class to be classified as the i^{th} class, $\sum_{i=1}^{k} X_{i|j} = n_j$, and $\sum_{i=1}^{k} p_{i|j} = 1$. Also, each observation can only be classified as one

Table 2.4: Contingency table for a k-class classification system. Columns represent truth and rows represent the label given by the classification system.

	Class 1	Class 2	Class 3	....	Class k
Test = 1	$X_{1\|1}$	$X_{1\|2}$	$X_{1\|3}$	....	$X_{1\|k}$
Test = 2	$X_{2\|1}$	$X_{2\|2}$	$X_{2\|3}$	....	$X_{2\|k}$
Test = 3	$X_{3\|1}$	$X_{3\|2}$	$X_{3\|3}$	....	$X_{3\|k}$
....	....	....	....	....	....
Test = k	$X_{k\|1}$	$X_{k\|2}$	$X_{k\|3}$	....	$X_{k\|k}$

outcome, resulting in $E[x_{i|j} \times x_{i'|j}, i \neq i'] = 0$. The multinomial pmf is

$$f_{\mathbf{X}}(\mathbf{x} \mid n, \mathbf{p}) = P(X_1 = x_1, X_2 = x_2, ..., X_k = x_k \mid n, \mathbf{p}) = \prod_{i=1}^{k} \frac{n!}{x_i!} p_i^{x_i}, \text{where } x_i \in (0, ..., n) \tag{2.45}$$

Each X_i considered individually (collapsing among the other $i - 1$ X's within class j) is distributed binomial(n, p_i). However, when considering all classification outcomes simultaneously, the multinomial distribution is used as it allows for consideration of multiple classification outcomes at once, and provides for the covariance structure between outcomes within a class. The MLEs of the multinomial parameters are

$$\widehat{\mathbf{p}} = (\widehat{p_1}, \widehat{p_2}, ..., \widehat{p_k}) = \left(\frac{x_1}{n}, \frac{x_2}{n}, ..., \frac{x_k}{n}\right) \tag{2.46}$$

where each x_i is the ith observed outcome and n is the total sample size [3, p.21].

2.7.2.1 Confidence Intervals for Multinomial Proportions.

In this section, methods available for simultaneous CIs for multinomial probabilities and linear combinations of multinomial probabilities are presented. In 1963, Gold introduced a CI for the linear combination of multinomial probabilities. Letting $\boldsymbol{l} = (l_1, \ldots, l_k)$ denote the linear multipliers for each probability:

$$\sum_{i=1}^{k} l_i p_i \in \sum_{i=1}^{k} l_i \widehat{p_i} \pm \left(\chi^2_{k-1,\alpha}\right)^{\frac{1}{2}} \left[\sum_{i=1}^{k} l_i^2 \widehat{p_i} - \left(\sum_{i=1}^{k} l_i \widehat{p_i}\right)^2\right]^{\frac{1}{2}} \left(\frac{1}{n}\right)^{\frac{1}{2}} \tag{2.47}$$

[26][53, p.217]. Gold also extended this for all linear combinations of several populations of multinomial probabilities, p_{ij} , as

$$\sum_{ij} l_{ij} p_{ij} \in \sum_{ij} l_{i,j} \widehat{p_{ij}} \pm (\chi^2_{r(k-1),\alpha})^{\frac{1}{2}} s_l \tag{2.48}$$

where j denotes the r populations ($j = 1, \ldots, r$), i denotes the c categories ($i = 1, \ldots, c$) and

$$s_l^2 = \sum_{j=1}^{r} \frac{1}{n_i} \left[\sum_{i=1}^{k} l_{ij}^2 \widehat{p_{ij}} - \left(\sum_{i=1}^{k} l_{ij} \widehat{p_{ij}} \right)^2 \right] \tag{2.49}$$

[53, p. 219]. When the linear combinations considered are contrasts, the degrees of freedom are reduced from $r(k-1)$ to $(r-1)(k-1)$, resulting in shorter intervals [27].

In 1964, Queensberry and Hurst found the solutions to the following quadratic equations

$$(\widehat{p_i} - p_i)^2 = \chi^2_{k-1,\alpha} \frac{p_i(1-p_i)}{n}, i = 1, \ldots, k \tag{2.50}$$

produced simultaneous CIs around multinomial probabilities [51][53, p.217].

Goodman (1965) used Bonferroni intervals where

$$p_i \in \widehat{p_i} \pm z_{\alpha/2k} \left[\frac{\widehat{p_i}(1-\widehat{p_i})}{n} \right]^{\frac{1}{2}} \tag{2.51}$$

[53, p.216][71]. This is equivalent to a Wald CI with a Bonferroni correction for multiple comparisons, but does not take into account the covariance between the multinomial parameters.

Fitzpatrick and Scott (1987) also introduced simultaneous CIs for multinomial parameters in [22] where,

$$p_i \in \widehat{p_i} \pm \frac{z_{\alpha/2}}{2\sqrt{n}} \tag{2.52}$$

All of these previous methods were developed with large sample theory.

Finally, in 1995 Sison and Glaz determined that a simultaneous CI for multinomial parameters can be found by first estimating the value of c where

$$v(c) = P(x_i - c \leq X_i^* \leq x_i + c; i = 1, \ldots, k) = 1 - \alpha \tag{2.53}$$

and X_i^* has a multinomial distribution with n and $\widehat{\mathbf{p}} = (\widehat{p_1}, \ldots, \widehat{p_k})$ [62]. Then define

$$\gamma = \frac{[(1-\alpha) - v(c)]}{[v(c+1) - v(c)]} \tag{2.54}$$

and the following skewed confidence region is recommended:

$$\left(\widehat{p}_i - \frac{c}{n} \leq p_i \leq \widehat{p}_i + \frac{(c + 2\gamma)}{n}; i = 1, \ldots, k\right) \tag{2.55}$$

[62]. Determining the CI in Equation 2.55 may be difficult, however this method is coded to be implemented in SAS software [39]. The SAS code was later adapted into the *MultinomialCI* package for R, which makes this CI very easy to use [52, 69].

2.8 Summary

Optimal points are important for classification systems, as they represent a system's optimal performance with respect to classification accuracy. Metrics for characterizing the performance of a classification system's optimal point are developed by the maximization of correct classification probabilities or minimization of misclassification probabilities (i.e. J and BC). Minimization of the misclassification probabilities allows for more flexibility in the optimal point selection, and therefore is chosen as a focus for this work.

Little work has been done previously to quantify the uncertainty around BC. Thus, methods for quantifying uncertainty in a classification system's BC value are derived and presented in the following chapters, for both parametric and nonparametric settings. Confidence intervals and hypothesis tests are developed to provide a range of flexible inference methods.

III. Parametric Confidence Intervals

3.1 Introduction

The purpose of this chapter is to derive CIs for BC, for any number of k classes, in order to quantify the optimal performance of a classification system and compare systems based upon performance criteria. These methods are developed under the assumption of a single feature that is independently and normally distributed for each class, because the feature used for the classification is often assumed to follow a continuous distribution, most commonly normal [23, 30, 33, 36, 40, 45, 47, 49, 54–56, 58, 64, 65, 75, 79, 80]. Placing a parametric assumption on the feature distributions allows for the use of convenient statistical methods for the evaluation of the classification system, with accurate results when the parametric assumptions are correct. Also, the assumption of a normally distributed feature is useful as often transformations to normality are common place when the feature follows a skewed continuous distribution, such as gamma or log-normal [45].

In Section 3.2, the delta method is used to approximate the variance of BC and the optimal thresholds for the development of their associated CIs. A numerical estimation technique is also presented as a method for efficiently estimating the partial derivatives that are required for the delta method approximations. Numerical estimation is especially useful (and necessary) when there are more than two classes, as it can be used to solve equations which are difficult or impossible to solve analytically, while remaining very accurate [25, p.1]. In fact, the optimal thresholds for BC must be found numerically (when weights on misclassification probabilities are not equal), and therefore, their partial derivatives with respect to the normal distribution parameters ($2k^2 - 2k$ of them) must also be solved numerically. Although the $2k$ partial derivatives of BC with respect to the normal distribution parameters can be found analytically, the derivation becomes cumbersome for large k. Therefore, numerical estimation techniques allow for easy extension of the delta method CIs to k classes.

In Section 3.3 GCIs are derived for the k-class J and BC, again assuming a single feature that is independently and normally distributed for each class. Although CIs for BC are the focus of this work, the GCI for the extended J is also presented as it is not currently available in the literature.

GCIs for the optimal thresholds are also presented. In Section 3.4, available bootstrap CI methods which may be used when the classification system is developed with parametric assumptions are discussed. Simulation results are presented in Section 3.5. The simulation examines the performance of the delta and generalized CIs, and compares these CIs' performance to that of available bootstrap CIs. Specifically, coverage probability, coverage symmetry, length of CIs and bias of $\widehat{BC}$ are assessed under a variety of classification system settings, including varying distributional parameters and costs. Finally, the results are summarized in Section 3.6.

3.2 Delta Method Confidence Intervals

The delta method uses the first order Taylor series expansion to estimate the variance of functions of parameters [12, p.242]. A multivariate version of the delta method is given in the following theorem.

Theorem 4 (Multivariate Delta Method)**.**

Suppose that $\widehat{\boldsymbol{\theta}}$ *is Asymptotic-Normal*$_k(\boldsymbol{\theta}, b_n^2\boldsymbol{\Sigma})$ *with* $b_n \to 0$ *and that* g *is a real-valued function with partial derivatives existing in a neighborhood of* $\boldsymbol{\theta}$ *and continuous at* $\boldsymbol{\theta}$ *with* $g'(\boldsymbol{\theta}) = \partial g(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ *not identically zero. Then as* $n \to \infty$

$$g(\widehat{\boldsymbol{\theta}}) \text{ is Asymptotic-Normal}[g(\boldsymbol{\theta}), b_n^2 g'(\boldsymbol{\theta})\boldsymbol{\Sigma} g'(\boldsymbol{\theta})^T]$$

[8, p. 238]

Often b_n is taken to be $\frac{1}{n}$ [8, p. 238]. In Theorem 4, $\boldsymbol{\theta}$ is used to represent any vector of statistical parameters. This theorem is applied for BC and the optimal threshold values, θ_m^* , which are both functions of $(\boldsymbol{\mu}, \boldsymbol{\sigma^2})$.

3.2.1 Bayes Cost and Optimal Thresholds, 3 classes.

Recall, if the classification system is developed using a single feature for the classification of three classes, where the feature is independently and normally distributed for each class with $\mu_1 < \mu_2 < \mu_3$ and with two threshold values used to distinguish between the classes (denoted $\theta_1 < \theta_2$), BC can be expressed using the standard normal CDF and the optimal threshold values

which minimize BC as:

$$BC_3 = c_{2|1}p_1 \times \left(\Phi\left(\frac{\theta_2^* - \mu_1}{\sigma_1}\right) - \Phi\left(\frac{\theta_1^* - \mu_1}{\sigma_1}\right)\right) + c_{3|1}p_1 \times \left(\Phi\left(\frac{\mu_1 - \theta_2^*}{\sigma_1}\right)\right)$$
$$+ c_{1|2}p_2 \times \left(\Phi\left(\frac{\theta_1^* - \mu_2}{\sigma_2}\right)\right) + c_{3|2}p_2 \times \left(\Phi\left(\frac{\mu_2 - \theta_2^*}{\sigma_2}\right)\right)$$
$$+ c_{1|3}p_3 \times \left(\Phi\left(\frac{\theta_1^* - \mu_3}{\sigma_3}\right)\right) + c_{2|3}p_3 \times \left(\Phi\left(\frac{\theta_2^* - \mu_3}{\sigma_3}\right) - \Phi\left(\frac{\theta_1^* - \mu_3}{\sigma_3}\right)\right) \quad (2.21)$$

Note that the minimization is not expressed in Equation 2.21 as it uses the optimal thresholds ($\theta_1^* < \theta_2^*$). The optimal thresholds must be estimated numerically when all $c_{i|j}p_j$ are unequal, for $i \neq j$. For BC defined in Equation 2.21, $\widehat{BC} = g(\overline{\boldsymbol{x}}, \boldsymbol{S^2})$. Since $(\overline{\boldsymbol{x}}, \boldsymbol{S^2})$ are asymptotically multivariate normal, the multivariate delta method may be applied (see Appendix A.1 for asymptotic properties of $(\overline{x}, S^2)$). Therefore, by Theorem 4, $\widehat{BC}$ is Asymptotic-Normal$[BC, Var(\widehat{BC})]$ and the variance of $\widehat{BC}$ from Equation 2.21 is estimated according to the delta method using the following equation:

$$Var(\widehat{BC_3}) \approx \left(\frac{\partial BC_3}{\partial \mu_1}\right)^2 Var(\widehat{\mu_1}) + \left(\frac{\partial BC_3}{\partial \mu_2}\right)^2 Var(\widehat{\mu_2}) + \left(\frac{\partial BC_3}{\partial \mu_3}\right)^2 Var(\widehat{\mu_3})$$
$$+ \left(\frac{\partial BC_3}{\partial \sigma_1}\right)^2 Var(\widehat{\sigma_1}) + \left(\frac{\partial BC_3}{\partial \sigma_2}\right)^2 Var(\widehat{\sigma_2}) + \left(\frac{\partial BC_3}{\partial \sigma_3}\right)^2 Var(\widehat{\sigma_3}) \quad (3.1)$$

where all covariances are zero due to the assumption of independence between the feature's distributions for each class. Letting $\widehat{\mu_j} = \overline{x}_j$ and $\widehat{\sigma_j} = S_j$, $Var(\widehat{\mu_j})$ and $Var(\widehat{\sigma_j})$ are [56]

$$Var(\widehat{\mu_j}) = \frac{\sigma_j^2}{n_j} \quad (3.2)$$

and

$$Var(\widehat{\sigma_j}) = \frac{\sigma_j^2}{2(n_j - 1)} \quad (3.3)$$

Thus, to estimate Equation 3.1, the partial derivative of BC_3 with respect to the normal distribution parameters, γ_j (where $\gamma_j = \mu_j$ or σ_j and $j = 1, 2, 3$), are defined

$$\frac{\partial BC_3}{\partial \gamma_j} = \frac{1}{\sigma_1}\left[\frac{\partial \theta_2^*}{\partial \gamma_j}A - \frac{\partial \theta_1^*}{\partial \gamma_j}B\right] + \frac{1}{\sigma_2}\left[\frac{\partial \theta_1^*}{\partial \gamma_j}C - \frac{\partial \theta_2^*}{\partial \gamma_j}D\right] + \frac{1}{\sigma_3}\left[\frac{\partial \theta_1^*}{\partial \gamma_j}E + \frac{\partial \theta_2^*}{\partial \gamma_j}F\right] + \delta_{\gamma_j} \quad (3.4)$$

where for $\gamma_j = \mu_j$

$$\delta_{\mu_j} = \begin{cases} \frac{1}{\sigma_1}(A - B) & \text{for j=1} \\ \frac{1}{\sigma_2}(B - C) & \text{for j=2} \\ -\frac{1}{\sigma_3}(E + F) & \text{for j=3} \end{cases}$$

or for $\gamma_j = \sigma_j$

$$\delta_{\sigma_j} = \begin{cases} B\left(\frac{\theta_1^* - \mu_1}{\sigma_1^2}\right) - A\left(\frac{\theta_2^* - \mu_1}{\sigma_1^2}\right) \text{ for j=1} \\ D\left(\frac{\theta_2^* - \mu_2}{\sigma_2^2}\right) - C\left(\frac{\theta_1^* - \mu_2}{\sigma_2^2}\right) \text{ for j=2} \\ E\left(\frac{\mu_3 - \theta_1^*}{\sigma_3^2}\right) - F\left(\frac{\mu_3 - \theta_2^*}{\sigma_3^2}\right) \text{ for j=3} \end{cases}$$

and for both μ_j and σ_j

$$A = p_1 \left(c_{2|1} - c_{3|1}\right) \phi\left(\frac{\theta_2^* - \mu_1}{\sigma_1}\right)$$

$$B = p_1 c_{2|1} \phi\left(\frac{\theta_1^* - \mu_1}{\sigma_1}\right)$$

$$C = p_2 c_{1|2} \phi\left(\frac{\theta_1^* - \mu_2}{\sigma_2}\right)$$

$$D = p_2 c_{3|2} \phi\left(\frac{\mu_2 - \theta_2^*}{\sigma_2}\right)$$

$$E = p_3 \left(c_{1|3} - c_{2|3}\right) \phi\left(\frac{\theta_1^* - \mu_3}{\sigma_3}\right)$$

$$F = p_3 c_{2|3} \phi\left(\frac{\theta_2^* - \mu_3}{\sigma_3}\right)$$

A more detailed derivation of these results is presented in the Appendix, Section A.2. The six partial derivatives of BC with respect to μ_j and σ_j in Equation 3.4 are estimated using the numerical estimates for $\frac{\partial \theta_m^*}{\partial \mu_j}$ and $\frac{\partial \theta_m^*}{\partial \sigma_j}$ (described in Section 3.2.3, $m = 1, 2,\ j = 1, 2, 3$) as well as $\widehat{\mu_j} = \bar{x}_j$ and $\widehat{\sigma_j} = S_j$. Using these estimated partial derivatives, the variance of $\widehat{BC_3}$ is estimated. The $(1 - \alpha)100\%$ delta method CI for the three-class BC is

$$\widehat{BC_3} \pm z_{\frac{\alpha}{2}} \sqrt{Var(\widehat{BC_3})} \tag{3.5}$$

Confidence intervals around the optimal thresholds in addition to the CI for BC are also of interest. There are three solutions for determining the optimal thresholds in this parametric framework. First, when all $c_{i|j} p_j$ are equal, for $i \neq j$, the optimal thresholds may be found equivalently as with J (see Section 2.5.2). Therefore, the solution for the optimal thresholds is

$$\theta_m^* = \frac{\mu_m(b^2 - 1) - a + b\sqrt{a^2 + 2(b^2 - 1)\sigma_1^2 \ln(b)}}{(b^2 - 1)} \tag{3.6}$$

where $a = \mu_{m+1} - \mu_m$ and $b = \frac{\sigma_{m+1}}{\sigma_m}$, $m = 1, \ldots, k-1$ [56]. Second, if $\sigma_m = \sigma_{m+1}$ the optimal point is the midpoint between the means [56]:

$$\theta_m^* = \frac{\mu_m + \mu_{m+1}}{2} \tag{3.7}$$

Finally, when all $c_{i|j}p_j$ are not equal, for $i \neq j$, the optimal thresholds must be estimated using numerical minimization (see Section 3.2.3). Whether θ_m^* is found using Equation 3.6, Equation 3.7, or numerically, the optimal thresholds' estimates are functions of the sample mean and variance $(\widehat{\theta_m^*} = f(\bar{\boldsymbol{x}}, \boldsymbol{S^2}))$. By Theorem 4, $\widehat{\theta_m^*}$ is Asymptotic-Normal$[\theta_m^*, Var(\widehat{\theta_m^*})]$, and the delta method approximate variance for each of the two optimal thresholds is given by

$$\begin{aligned} Var\left(\widehat{\theta_m^*}\right) \approx \left(\frac{\partial \theta_m^*}{\partial \mu_1}\right)^2 Var(\widehat{\mu_1}) + \left(\frac{\partial \theta_m^*}{\partial \mu_2}\right)^2 Var(\widehat{\mu_2}) + \left(\frac{\partial \theta_m^*}{\partial \mu_3}\right)^2 Var(\widehat{\mu_3}) \\ + \left(\frac{\partial \theta_m^*}{\partial \sigma_1}\right)^2 Var(\widehat{\sigma_1}) + \left(\frac{\partial \theta_m^*}{\partial \sigma_2}\right)^2 Var(\widehat{\sigma_2}) + \left(\frac{\partial \theta_m^*}{\partial \sigma_3}\right)^2 Var(\widehat{\sigma_3}) \end{aligned} \tag{3.8}$$

This estimate provides a $(1-\alpha)100\%$ delta method CI for each optimal threshold of $\widehat{\theta_m^*} \pm z_{\frac{\alpha}{2}} \sqrt{Var(\widehat{\theta_m^*})}$, as was demonstrated in [65].

3.2.2 Bayes Cost and Optimal Thresholds, k classes.

These methods extend easily for $k > 3$ classes. When there are k classes, BC may be expressed using the normal CDF as

$$\begin{aligned} BC = \sum_{j=2}^{k} c_{1|j} p_j \Phi\left(\frac{\theta_1^* - \mu_j}{\sigma_j}\right) \\ + \sum_{\substack{i=2 \\ i \neq j}}^{k-1} \sum_{j=1}^{k} c_{i|j} p_j \left[\Phi\left(\frac{\theta_{m=i}^* - \mu_j}{\sigma_j}\right) - \Phi\left(\frac{\theta_{m=i-1}^* - \mu_j}{\sigma_j}\right)\right] + \sum_{j=1}^{k-1} c_{k|j} p_j \Phi\left(\frac{\mu_j - \theta_{k-1}^*}{\sigma_j}\right) \end{aligned} \tag{3.9}$$

The $(1-\alpha)100\%$ CI for BC is still $\widehat{BC} \pm z_{\frac{\alpha}{2}} \sqrt{Var(\widehat{BC})}$ where

$$Var\left(\widehat{BC}\right) \approx \sum_{j=1}^{k} \left[\left(\frac{\partial BC}{\partial \mu_j}\right)^2 Var(\widehat{\mu_j}) + \left(\frac{\partial BC}{\partial \sigma_j}\right)^2 Var(\widehat{\sigma_j})\right] \tag{3.10}$$

and the partial derivatives may be estimated using Equation 3.4 for three classes, Equations A.4 through A.11 in Appendix A.3 for four classes, or the methods described in Section 3.2.3 below for any k classes. Similarly, the $(1-\alpha)100\%$ CI for each optimal threshold is $\widehat{\theta_m^*} \pm z_{\frac{\alpha}{2}} \sqrt{\widehat{Var}(\widehat{\theta_m^*})}$ where

$$Var\left(\widehat{\theta_m^*}\right) \approx \sum_{j=1}^{k} \left[\left(\frac{\partial \theta_m^*}{\partial \mu_j}\right)^2 Var(\widehat{\mu_j}) + \left(\frac{\partial \theta_m^*}{\partial \sigma_j}\right)^2 Var(\widehat{\sigma_j})\right] \tag{3.11}$$

When all $c_{i|j}p_j$ are equal (WLOG assumed to be one), for $i \neq j$, the partial derivatives are given in Equations 2.28 through 2.31 with class 1 and class 2 being replaced with class $j = m$ and class $j = m + 1$ for the m^{th} optimal threshold ($m = 1, \ldots, k - 1$). When the costs and prevalences are not equal, the optimal thresholds must be found numerically and the partial derivatives are estimated using Equation 3.13. Note that these CIs for k classes define the CIs for $k = 2$ and 3 classes as well.

Finally, it is worth noting that there exists covariance between each m and $m + 1$ threshold, due to the thresholds' shared dependence on the feature's parameters of the class between them and may be estimated with the delta method as

$$Cov\left(\widehat{\theta}^*_m, \widehat{\theta}^*_{m+1}\right) = \left(\frac{\partial\theta^*_m}{\partial\mu_{j=m+1}}\right)\left(\frac{\partial\theta^*_{m+1}}{\partial\mu_{j=m+1}}\right)Var(\widehat{\mu}_{j=m+1}) + \left(\frac{\partial\theta^*_m}{\partial\sigma_{j=m+1}}\right)\left(\frac{\partial\theta^*_{m+1}}{\partial\sigma_{j=m+1}}\right)Var(\widehat{\sigma}_{j=m+1}) \quad (3.12)$$

[36]. This covariance may be used for constructing confidence regions around pairs of optimal thresholds.

3.2.3 A Method for Numerically Estimating Partial Derivatives.

Although the solutions to the optimal thresholds, θ^*_m , are functions of the distributional parameters, they generally must be found numerically when minimizing BC. Therefore, the partial derivatives, $\frac{\partial\theta^*_m}{\partial\mu_j}$ and $\frac{\partial\theta^*_m}{\partial\sigma_j}$ ($j = 1, ..., k$ represents the true class and $m = 1, ..., k-1$ denotes the optimal thresholds), must also be estimated numerically. This can be accomplished using the two-point central difference method [25, p. 254]. Applying this method, for $\gamma_j = \mu_j$ or σ_j

$$\frac{\partial\theta^*_m}{\partial\gamma_j} \approx \frac{\theta^*_m(\gamma_j + \varepsilon) - \theta^*_m(\gamma_j - \varepsilon)}{2\varepsilon} \quad (3.13)$$

leaving all other normal parameters constant for each calculation. The term $\theta^*_m(\gamma_j \pm \varepsilon)$ is determined using the same numerical minimization method as that used to find the optimal threshold values. The truncation error for this difference method is $O(\varepsilon^2)$. The ε value should be chosen to minimize the error of the approximation, which for double precision (using 64 bits to store values) would be

$$\text{Error} \approx \frac{10^{-16}}{\varepsilon} + O(\varepsilon^2) \quad (3.14)$$

This error would be minimized for ε on the order of $10^{-16/3}$. Therefore, a small ε should be chosen; however, ε should be $\geq 10^{-5}$ to avoid inflating the error caused by computer precision.

The partial derivatives of BC can be found analytically, and were presented in Section 3.2.1 for three classes and in the Appendix, Section A.3 for four classes. For any k classes, the partial derivatives for BC can be approximated by

$$\frac{\partial BC}{\partial \gamma_j} \approx \frac{BC(\gamma_j + \varepsilon) - BC(\gamma_j - \varepsilon)}{2\varepsilon} \tag{3.15}$$

where $BC(\gamma_j \pm \varepsilon)$ is found using Equation 3.9 for $\gamma_j = \mu_j$ or σ_j , and using equivalent values for ε as discussed for θ_m^* .

3.3 Generalized Confidence Intervals

In [33], GCIs[2] are developed for the two-class J as an exact method for constructing CIs around J and the optimal threshold when the feature used for classification is independently and normally distributed for each class. Define $\zeta = (\theta, \boldsymbol{\delta})$ where θ is the parameter of interest and $\boldsymbol{\delta}$ is a vector of nuisance parameters.

Definition 1 (Generalized Pivotal Quantity)**.**

> *Let $R = r(\mathbf{X}; \mathbf{x}, \zeta)$ be a function of $\mathbf{X}$ and possibly $\mathbf{x}$, ζ as well. The random quantity R is said to be a generalized pivotal quantity if it has the following two properties:*
> ***Property A**: R has a probability distribution that is free of unknown parameters.*
> ***Property B**: r_{obs} defined as $r_{obs} = r(\mathbf{x}; \mathbf{x}, \zeta)$ ... does not depend on nuisance parameters, δ. [73, p. 146]*

Definition 2 (Generalized Confidence Interval)**.**

> *If the subset C_γ of the sample space ρ of R satisfies ($Pr(R \in C_\gamma) = \gamma$), then the subset Θ_C of the parameter space given by $\Theta_C(r) = \{\theta \in \Theta \mid r(\mathbf{x}; \mathbf{x}, \zeta) \in C_\gamma\}$ is said to be a $100\gamma\%$ GCI for θ. [73, p. 146]*

3.3.1 *Youden Index, k Classes.*

In [33], a GCI is developed for the two-class J by constructing generalized pivotal quantities (GPQs) for μ_j and σ_j ($j = 1, 2$), and then using these pivotal quantities to construct GPQs for the optimal threshold and J. For a classification system with k classes and a normally distributed feature, there will be $k - 1$ optimal thresholds, one between each pair of normal distributions. Therefore, in order to extend this method for k classes, $k - 1$ GPQs for the optimal thresholds must be determined and used to define the GPQ for J (defined as the sum of all correct classification rates). Each optimal

[2] In [66] it was noted that the implementation of these GCIs is identical to constructing the CIs via Bayesian Inference using the non-informative prior ($p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$).

threshold value is determined explicitly by the distributions of the two classes it divides [46]. For this problem, J is the parameter of interest and the mean (μ_j) and variance (σ_j^2) from each class are the nuisance parameters. Then (as is done in [33] for two classes) define

$$R_{\mu_j} = \bar{x}_j - t_j \frac{S_j}{\sqrt{n_j}} \tag{3.16}$$

$$R_{\sigma_j} = \sqrt{\frac{(n_j - 1)S_j^2}{V_j}} \tag{3.17}$$

where

$$t_j = \frac{\overline{X}_j - \mu_j}{S_j/\sqrt{n_j}} \tag{3.18}$$

and

$$V_j = \frac{(n_j - 1)S_j^2}{\sigma_j^2} \tag{3.19}$$

The sample mean ($\bar{x}_j$) and standard deviation (S_j) are from the j^{th} class, $t_j \sim t_{(n_j-1)}$, a t-distribution random variable with $n_j - 1$ degrees of freedom, and $V_j \sim \chi^2_{n_j-1}$, a chi-square random variable with $n_j - 1$ degrees of freedom [12, p. 218, 223]. To find the $k-1$ GPQs for the optimal thresholds ($R_{\theta_m^*}$, indexed on $m = 1, 2, \ldots, k-1$), first define the following $k-1$ GPQs

$$R_{a_m} = R_{\mu_{j=m+1}} - R_{\mu_{j=m}} \tag{3.20}$$

$$R_{b_m} = \frac{R_{\sigma_{j=m+1}}}{R_{\sigma_{j=m}}} \tag{3.21}$$

Next, the GPQs for the $k-1$ optimal thresholds are computed as

$$R_{\theta_m^*} = \frac{R_{\mu_{j=m}}(R_{b_m}^2 - 1) - R_{a_m} + R_{b_m}\sqrt{R_{a_m}^2 + (R_{b_m}^2 - 1)R_{\sigma_{j=m}}\ln(R_{b_m}^2)}}{R_{b_m}^2 - 1} \tag{3.22}$$

for $m = 1, 2, \ldots, k-1$. Using these GPQs, the GPQ for the k-class J is defined as

$$R_J = \Phi\left(\frac{R_{\theta_1^*} - R_{\mu_1}}{R_{\sigma_1}}\right) + \sum_{j=2}^{k-1}\left[\Phi\left(\frac{R_{\theta_{m=j}^*} - R_{\mu_j}}{R_{\sigma_j}}\right) - \Phi\left(\frac{R_{\theta_{m=j-1}^*} - R_{\mu_j}}{R_{\sigma_j}}\right)\right] + \Phi\left(\frac{R_{\mu_k} - R_{\theta_{k-1}^*}}{R_{\sigma_k}}\right) \tag{3.23}$$

It is clear that R_{μ_j} and R_{σ_j} do not depend on any unknown parameters and therefore, $R_{\theta_m^*}$ and R_J (defined only with R_{μ_j} and R_{σ_j}) do not depend on unknown parameters. This satisfies property A of Definition 1. Also note, $r_{J_{obs}} = R_J(\bar{\mathbf{x}}, \mathbf{S})$ is evaluated by using $\bar{x}_j$ and S_j in Equations 3.18 and 3.19 and then substituting Equations 3.18 and 3.19 into Equations 3.16 and 3.17, respectively. This

results in $R_{\mu_j}(\bar{\mathbf{x}}, \mathbf{S}) = \mu_j$, $R_{\sigma_j}(\bar{\mathbf{x}}, \mathbf{S}) = \sigma_j$, and $R_{\theta_m^*}(\bar{\mathbf{x}}, \mathbf{S}) = \theta_m^*$. Evaluating R_J with these values gives $R_J(\bar{\mathbf{x}}, \mathbf{S}) = J$; therefore $r_{J_{obs}}$ does not depend on any nuisance parameters and property B of Definition 1 is met.

Finally, a CI around J can be found using Monte Carlo simulation by generating a large number (K$\approx$ 2,500) of random draws from t_j and V_j for each class, $j = 1, \ldots, k$. Using these values in Equations 3.16 through 3.23, K R_J values are calculated. Then the $(\frac{\alpha}{2})100^{th}$ and $(1 - \frac{\alpha}{2})100^{th}$ percentiles of R_J are defined to be the lower and upper bounds for the $(1 - \alpha)100\%$ GCI around J, respectively [33]. Also note, the $(1 - \alpha)100\%$ GCI around each of the $k - 1$ optimal thresholds can be found similarly using the appropriate percentiles of each $R_{\theta_m^*}$ GPQ ($m = 1, \ldots, k - 1$).

3.3.2 Bayes Cost, Equal Weights.

The GCI around BC from a classification system with equal $c_{i|j}p_j$, for $i \neq j$ (WLOG $p_j c_{i|j} = 1$, accomplished by scaling BC by the reciprocal of the common multiplier), are found using the GPQs for the mean, standard deviation, and $k - 1$ optimal thresholds in Equations 3.16 - 3.22. The mean and variance of the feature's distribution for each class are still the nuisance parameters, and BC is the parameter of interest. Then the GPQ for BC is

$$R_{BC} = \sum_{j=2}^{k} \Phi\left(\frac{R_{\theta_1^*} - R_{\mu_j}}{R_{\sigma_j}}\right) + \sum_{\substack{i=2 \\ i \neq j}}^{k-1} \sum_{j=1}^{k} \left[\Phi\left(\frac{R_{\theta_{m=i}^*} - R_{\mu_j}}{R_{\sigma_j}}\right) - \Phi\left(\frac{R_{\theta_{m=i-1}^*} - R_{\mu_j}}{R_{\sigma_j}}\right)\right] + \sum_{j=1}^{k-1} \Phi\left(\frac{R_{\mu_j} - R_{\theta_{k-1}^*}}{R_{\sigma_j}}\right) \tag{3.24}$$

It is clear, as was discussed for R_J in the previous section, that R_{BC} is a GPQ meeting both properties of Definition 1. The $(1 - \alpha)100\%$ GCIs around BC and the optimal thresholds may be found using Monte Carlo simulation as was described for J and the optimal thresholds in Section 3.3.1.

3.3.3 Bayes Cost, Unequal Weights.

In this section, the GCI for BC from a classification system with unequal $c_{i|j}p_j$, for $i \neq j$, is developed. Once again, the nuisance parameters are the mean and variance of the feature's distributions for each class, and the parameter of interest is BC. With unequal costs, the GPQs for the optimal thresholds can no longer be found using the closed form solution in Equation 3.22. Although there is no closed form solution for $R_{\theta_m^*}$, the optimal thresholds are functions of the mean

and variance of each class and can be found with numerical minimization. The GPQ for BC is defined, now with costs and prevalences on the misclassification probabilities:

$$R_{BC} = \sum_{j=2}^{k} c_{1|j} p_j \Phi\left(\frac{R_{\theta_1^*} - R_{\mu_j}}{R_{\sigma_j}}\right) + \sum_{\substack{i=2 \\ i\neq j}}^{k-1} \sum_{j=1}^{k} c_{i|j} p_j \left[\Phi\left(\frac{R_{\theta^*_{m=i}} - R_{\mu_j}}{R_{\sigma_j}}\right) - \Phi\left(\frac{R_{\theta^*_{m=i-1}} - R_{\mu_j}}{R_{\sigma_j}}\right)\right] + \sum_{j=1}^{k-1} c_{k|j} p_j \Phi\left(\frac{R_{\mu_j} - R_{\theta^*_{k-1}}}{R_{\sigma_j}}\right) \quad (3.25)$$

The $k-1$ optimal threshold values' GPQs ($R_{\theta_m^*}$) are found numerically for each of the K sets of R_{μ_j} and R_{σ_j} values from Equations 3.16 through 3.19 (this requires K numerical minimizations of Equation 3.25, resulting in K $R_{\theta_m^*}$ values and K R_{BC} values). Once again, $R_{\theta_m^*} = f(\mathbf{R}_\mu, \mathbf{R}_\sigma)$ and each $R_{\theta_m^*}$ does not depend on any unknown parameters. Therefore, as was seen for J and BC with equal weights in Sections 3.3.1 and 3.3.2, R_{BC} in Equation 3.25 does not depend on unknown parameters and achieves property A of Definition 1. Also, $R_{\mu_j}(\bar{\mathbf{x}}, \mathbf{S}) = \mu_j$, $R_{\sigma_j}(\bar{\mathbf{x}}, \mathbf{S}) = \sigma_j$, and $R_{\theta_m^*}(\bar{\mathbf{x}}, \mathbf{S}) = \theta_m^*$, resulting in $r_{BC_{obs}} = R_{BC}(\bar{\mathbf{x}}, \mathbf{S}) = BC$, which does not depend on nuisance parameters. This satisfies property B of Definition 1. Once again, by randomly generating K values of t_j and V_j for each class, K R_{BC} and $R_{\theta_m^*}$ values are determined with numerical minimization. Then, the $(1-\alpha)100\%$ GCI around BC is determined as the $(\frac{\alpha}{2})100^{th}$ and $(1-\frac{\alpha}{2})100^{th}$ percentiles of R_{BC} (or similarly, the analogous percentiles of $R_{\theta_m^*}$ are used to construct GCIs around the optimal thresholds).

3.4 Bootstrap Methods

Bootstrap methods were introduced by Efron in the 1970s [12, p. 478]. The bootstrap can be used for creating CIs for large or small data samples where the assumptions inherent for other methods are not met. With increasing computing power, the bootstrap has become a popular method for constructing CIs. Typically, a nonparametric bootstrap sample $\mathbf{X}^* = (X_1^*, \ldots, X_n^*)$ is created from a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ where a new sample of size n is drawn from $\mathbf{X}$ with replacement. It is also possible to draw a parametric bootstrap sample, where an underlying distribution ($F_X(x \mid \widehat{\theta})$) is assumed known and where $\widehat{\theta}$ (or $\widehat{\boldsymbol{\theta}}$) is an estimate for the true parameter θ (or parameters, $\boldsymbol{\theta}$) from the initial sample $\mathbf{X}$ [11]. Then, the bootstrap sample $\mathbf{X}^*$ is created by sampling n times from the distribution ($F_X(x \mid \widehat{\theta})$). The work in this dissertation utilizes

nonparametric resampling, which is the sampling procedure most commonly used. Generally, a large number (B) of bootstrap samples ($\mathbf{X}^*$) are drawn in order to construct CIs.

One common bootstrap CI assumes asymptotic normality of the parameter estimate. This is accomplished by estimating the variance of the parameter estimate from the B bootstrap samples and using this variance with standard normal quantiles to construct a CI. This method generates what is known as an asymptotic normal (AN) bootstrap CI [14]. This method, however, is not robust under transformations of the parameters, and could also possibly include values in the interval that are not valid (for example, *BC* values less than zero) [11, 14]. Therefore, two other bootstrap CI methods are considered which are the basic percentile (BP) bootstrap CI and the bias corrected and accelerated (BCa) bootstrap CI. The advantage of the BP CI is that the resulting interval will not include invalid values of the parameter of interest, since the CI bounds are found as the appropriate percentiles from the B bootstrap estimates of the parameter. However, a disadvantage to this method is that the coverage will be low when the distribution of the estimated parameter is not symmetric [11]. The BCa CI has the same advantage of the BP CI, however also performs well for skewed distributions of the estimated parameter [11]. All three of these bootstrap CIs are implemented using the *boot.ci* function in the *boot* package in R [10, 15, 52]. For more information on the bootstrap see [15].

The performance of the bootstrap CIs may be impacted by the method used for estimation of the parameter of interest, as different estimation techniques result in different levels of bias depending on the true scenario (here, classification system structure as well as feature distributions). For comparison to the parametric CIs for *BC* presented in this chapter, the point estimates for *BC* and θ^* are estimated parametrically as is done, for example, in Equations 2.21 and 3.6, respectively.

3.5 Simulation Results

A simulation study was conducted to demonstrate the performance of the delta method and generalized CIs around *BC*, and compare their performance to available bootstrap CI methods. The performance of CIs around the optimal thresholds is also evaluated. Various classification scenarios are considered including different sample sizes, underlying distributions of the feature used for classification, differing costs associated with the misclassifications, and classification accuracy

(measured by the BC value). All scenarios assume a classifier with three classes and two optimal thresholds ($\theta_1^* < \theta_2^*$) to distinguish between adjacent classes. Thus BC_3 could range from completely accurate, $BC_3 = 0.0$, to misclassifying all observations, $BC_3 = 3.0$. Five BC_3 values are chosen to demonstrate a range of classification system performances (all better than chance classification which occurs for $BC_3 = 1.5$). These values are $BC_3 = 0.27$, 0.42, 0.63, 0.91, and 1.23. The distributional parameters for each class are determined by varying each distribution's mean and variance in order to achieve the desired BC_3 value. The parameters for all scenarios are presented in Table 3.1.

In Section 3.5.1, it is assumed that all $c_{i|j}p_j$ are equal, for $i \neq j$. Using this equal cost/prevalence structure, various distributions on the feature are considered in order to study the impact of non-normal distributions on the performance of the CI methods. Therefore, the CIs are applied as described in this chapter, using the methods derived for normally distributed features. In Section 3.5.2, two additional cost structures are used to determine if unequal cost scenarios alter the CIs' performance. These different costs are applied to the same normal distribution settings in Table 3.1 ($\sigma_3 = 1$), however the resulting BC_3 values change due to the multiplication of the different costs on the misclassification probabilities. Although the normal distributions are unchanged, the different cost structures also result in different optimal thresholds between the classes, as is expected when accounting for the costs placed on the different classification errors.

The bootstrap CI methods considered for comparison are the BP, AN, and BCa. All bootstrap CIs utilize 1,000 nonparametric resamples and estimate BC_3 parametrically (Equation 2.21). The optimal thresholds ($\theta_1^* < \theta_2^*$) are found with Equation 3.6 or via numerical minimization for each resample, for equal and unequal costs respectively. Equation 2.21 is also used to estimate BC_3 for the delta method CIs. A similar parametric formulation is used for the GCIs, eliminating the impact of bias on comparisons of coverage probability between the different CI methods. Random samples from each of the three classes are generated of sizes n_j =10 to 250 from the appropriate distributions for all scenarios. This is repeated 5000 times (3000 times for the GCIs due to computational time) to determine the coverage probability, left and right coverage probability (for CI symmetry), and

Table 3.1: Distributional parameters for the parametric CI simulation.

Distribution	BC_3	Class 1		Class 2		Class 3	
Normal		μ	σ	μ	σ	μ	σ
($\sigma_3 = 1$)	1.23	-1	1	0	1	1	1
	0.91	-1.5	1	0	1	1.5	1
	0.63	-2	1	0	1	2	1
	0.42	-2.5	1	0	1	2.5	1
	0.27	-3	1	0	1	3	1
Normal		μ	σ	μ	σ	μ	σ
($\sigma_3 = 2$)	1.23	-1	1	0	1	1.2	2
	0.91	-1.5	1	0	1	2	2
	0.63	-2	1	0	1	2.85	2
	0.42	-2.5	1	0	1	3.6	2
	0.27	-3	1	0	1	4.4	2
Normal		μ	σ	μ	σ	μ	σ
($\sigma_3 = 4$)	1.23	-1	1	0	1	1	4
	0.91	-1.5	1	0	1	2.6	4
	0.63	-2	1	0	1	4.2	4
	0.42	-2.5	1	0	1	5.5	4
	0.27	-3	1	0	1	6.9	4
Gamma		α	β	α	β	α	β
	1.23	1.3	1	2	1.5	3	1.738
	0.91	1.3	1	2	1.5	3	3.544
	0.63	1.3	1	2	1.5	5	5.340
	0.42	1.3	1	2.3	3.7	5	6.463
	0.27	1.3	1	2.3	3.7	5	13.696
Normal		$\frac{1}{2}N(\mu,\sigma)$	$\frac{1}{2}N(\mu,\sigma)$	μ	σ	$\frac{1}{2}N(\mu,\sigma)$	$\frac{1}{2}N(\mu,\sigma)$
Mixtures	1.23	$\frac{1}{2}N(-1,2)$	$\frac{1}{2}N(-.2988,1)$	0	1	$\frac{1}{2}N(.800,1)$	$\frac{1}{2}N(3.600,1)$
	0.91	$\frac{1}{2}N(-2.235,1)$	$\frac{1}{2}N(-1,2)$	0	1	$\frac{1}{2}N(.800,1)$	$\frac{1}{2}N(3.600,1)$
	0.63	$\frac{1}{2}N(-4.5,1)$	$\frac{1}{2}N(-2,2)$	0	1	$\frac{1}{2}N(1.200,1)$	$\frac{1}{2}N(3.600,1)$
	0.42	$\frac{1}{2}N(-4.5,1)$	$\frac{1}{2}N(-2,2)$	0	1	$\frac{1}{2}N(2.417,1)$	$\frac{1}{2}N(4.817,1)$
	0.27	$\frac{1}{2}N(-4.5,1)$	$\frac{1}{2}N(-2,2)$	0	1	$\frac{1}{2}N(5.210,1)$	$\frac{1}{2}N(7.610,1)$

the average CI length. Absolute bias of the point estimates is also determined and is discussed throughout the following sections.

All simulations are run in R utilizing the *boot* package, and numerical minimization of *BC* is performed using the *optim* function with method "L-BFGS-B" [10, 15, 52]. The partial derivatives

of the optimal thresholds with respect to the normal distribution parameters are found numerically as described in Section 3.2.3 with the same *optim* function. The partial derivatives of BC with respect to the normal distribution parameters are calculated with Equation 3.4. Due to the large number of numeric results for this simulation, the tables of results are in the Appendix, Section B.1. A summary of these results follow.

3.5.1 Equal Costs and Prevalences.

All costs and prevalences are assumed equal, with a multiplier on each misclassification probability of one (i.e., $c_{i \neq j|j} = 3$, $p_j = \frac{1}{3}$). Four different feature distributions are simulated (normal, gamma, gamma transformed to normal (via Box-Cox), and normal mixtures). In addition, three normal distribution scenarios are considered, one with all $\sigma_j = 1$, one with $\sigma_1 = \sigma_2 = 1$ and $\sigma_3 = 2$, and one with $\sigma_1 = \sigma_2 = 1$ and $\sigma_3 = 4$.

3.5.1.1 Performance of Confidence Intervals around Bayes Cost.

The coverage probability and length for the delta, generalized, and bootstrap CIs when all $c_{i|j}p_j$ are assumed equal, for $i \neq j$, are presented in Table B.1 for a feature with independent normal distributions for each class and in Table B.2 for when the feature is not distributed normal. In general, the delta method, generalized, and bootstrapped BCa CIs perform similarly and better than the other two bootstrap CIs for BC_3. When the feature is normally distributed (equal or unequal variances), the length of all intervals are similar for $n_j \geq 50$ and the length of the delta method and generalized CIs are slightly larger than the bootstrap CIs for $n_j = 10$. However, the delta method CI performs slightly better than the BCa CI when considering coverage for $n_j = 10$ and the generalized CI performs the best with regards to coverage for $n_j = 10$ (only method to achieve coverage of at least 95%). For $n_j \geq 50$ both the delta method and BCa CIs have similar, good coverage ($\approx$ 93 - 95%). The GCI has better coverage than the delta method and BCa bootstrapped CIs for all sample sizes ($\approx$ 95 - 96%), with comparable lengths. Changing the value of σ_3 does not have a significant impact on the coverage for any of the methods.

The symmetry of the CIs for the normally distributed features are presented in Figure 3.1 for the delta method CI and Figure 3.2 for the GCI, with $\sigma_3 = 1$ and $\sigma_3 = 4$ in rows 1 and 2, respectively. The delta method CIs around BC_3 for both scenarios are skewed left, with the skew becoming less

extreme as n_j increases. The GCIs demonstrate an opposite trend in skewness, although notably much less extreme than that of the delta method (Right - Left coverage $\in$ [-.04,0.04] compared to Right - Left coverage $\in$ [-10,0] for the delta method). The $\widehat{BC_3}$ bias across all scenarios is low for the normally distributed features, as expected (absolute bias $\in$ $[.00003, .05]$). In general, the absolute bias decreases as n_j increases and increases when the BC_3 value increases (less accurate classification).

When the feature used for classification is distributed with an independent gamma for each class and is not transformed to normality, coverage probability for all CI methods is greatly diminished (see Table B.2). For all sample sizes in this scenario, the delta method and generalized CIs perform better than the BCa CI for accurate tests (BC_3=0.27 and 0.42), worse than BCa for very inaccurate tests (BC_3=1.23), and similar to the BCa CI for the other two scenarios. The one exception is for $n_j = 10$, where the GCI method performs better than the delta and BCa CIs. The bias of the estimates for the gamma distributed feature is slightly worse than with the normal distributed feature (absolute bias $\in$ $[.001, .09]$) and follow the same trend as the normal feature with respect to n_j and BC_3 values.

When the feature is distributed gamma and transformed to normality, the coverage probability is improved (Table B.2). However, overall, the coverage is slightly worse than when the feature is distributed normal, especially for the accurate scenarios ($BC_3 = 0.27$). The GCI has a slight advantage in coverage for this distributional scenario, although this results in longer intervals than the delta and BCa CIs. The bias of the estimates for BC_3 is very similar to that from normally distributed features (absolute bias $\in$ $[.00002, .06]$) and again has similar trends with n_j and BC_3. Finally, when the feature is distributed as independent normal mixtures for each class, the coverage probability for all methods is sporadic and poor, with the BCa CIs performing slightly better than the other methods (Table B.2). The bias of $\widehat{BC_3}$ for these distributions also represents the worst of all scenarios considered (absolute bias $\in$ $[.001, .13]$), with only slight improvements in bias for increases in n_j and decreases in BC_3 value.

3.5.1.2 Performance of Confidence Intervals around Optimal Thresholds.

The coverage probability and length of the delta, generalized, and all bootstrap CIs when all costs are assumed equal with a normally distributed feature are presented in Table B.3 for θ_1^* and Table B.4 for θ_2^* . Both the delta method and generalized CIs perform well with regards to coverage for both θ_1^* and θ_2^* ($\approx 91 - 97\%$). The GCI is the only method that achieves or exceeds the desired coverage of 95% for $n_j = 10$, however achieving this coverage results in CI lengths which are slightly longer than the other methods. For $n_j \geq 50$, the delta method, generalized, and AN bootstrap CIs perform similarly. Over all sample sizes when the variances are equal ($\sigma_3 = 1$), the GCIs have the best coverage and are only slightly longer in some scenarios.

When the variances are not equal ($\sigma_3 = 2$ or 4), the coverage and lengths of all CI methods are unchanged from the equal variance scenario for θ_1^* . However, θ_2^* depends on the third class's distributional parameters and therefore, the AN bootstrap CI does worse with respect to coverage around θ_2^* for $\sigma_3 = 2$ or 4. The delta method's coverage and lengths remain the same, and the GCI's performance also remains fairly constant. The BP and BCa bootstrap CIs have similar and better performance than the AN bootstrap CI when $\sigma_3 \neq 1$.

The bias of both optimal threshold estimates are equally good (absolute bias $\in [.00006, .03]$) when the variances are equal. The change in variance structure has no impact on the bias of θ_1^* , however the maximum absolute bias for θ_2^* increases from 0.02 to 0.05 when the variance of the third class changes. Symmetry is plotted for the delta method CIs (Figure 3.1) and the GCIs (Figure 3.2) around both optimal thresholds for $\sigma_3 = 1$ and $\sigma_3 = 4$, rows 1 and 2, respectively. For larger values of σ_3 , the symmetry of the delta method CI around θ_2^* becomes left skewed (row 2, Figure 3.1). Once again the GCI is less skewed than the delta method CI, and although the increase in σ_3 appears to have a slight impact on the symmetry of the GCI around θ_2^* , this change is very small compared to that seen with the delta method CI.

When the feature's distribution for each class is an independent gamma, the performance of all CI methods for θ_1^* and θ_2^* is extremely poor, and becomes worse as n_j increases (Tables B.5 and B.6 for θ_1^* and θ_2^* , respectively). Although using a Box-Cox transformation provides a slight increase in performance for all methods (Tables B.5 and B.6), the performance is still poor and

coverage is sporadic. When the Box-Cox transformation is used on the gamma distributions, the GCI and AN bootstrapped CI have a slight advantage with respect to coverage for most scenarios, although this advantage is minimal. The bias for the estimates of the optimal thresholds for the gamma distributions and the transformed gamma distributions is also poor. For the untransformed gamma distributions, the absolute bias of θ_1^* ranges from .01 to .9 and is largest for BC_3 values of 0.27 and 0.42. Additionally, the absolute bias of θ_2^* ranges from .0009 to 1.94 and performs best for BC_3 values of 0.63. The absolute bias increases as n_j increases for both optimal thresholds. For the transformed gamma distributions, the absolute bias of θ_1^* ranges from .02 to 4.65 and the absolute bias of θ_2^* ranges from .04 to 1.4, demonstrating worse estimation than the untransformed gamma distributions. Both of the threshold estimates have the largest bias for $n_j = 10$, and have similar bias for $n_j \geq 50$ (absolute bias $\in [.02, .09]$).

All CI methods have higher coverage with the normal mixtures than for both gamma scenarios for θ_2^* , but not for θ_1^* (coverage probability for θ_1^* with the normal mixtures is very poor). The absolute bias of θ_1^* ranges from .04 to .17. Again, bias increases as n_j increases. Also, the bias is lowest for BC_3 values of 0.91 and 1.23, which also corresponds to the best coverage for θ_1^* . The absolute bias of θ_2^* ranges from .000004 to .2, and as n_j increases the bias decreases. Also, as the BC_3 value decreases, the bias increases as does the coverage probability. The normal mixture distributions for the third class are mixes of normals with the same variance (equal to one) and different means. The normal mixtures for the first class have both different variances and means. Therefore, the shape of the normal mixture will have an impact on the performance of the CI around the threshold. The CI around the threshold associated with the mixture having equal variances performed fairly well when compared to the threshold adjacent to the mixtures with different variances.

3.5.2 Unequal Costs.

The unique advantage of using BC is the ability to consider different cost structures on the misclassification outcomes. The two cost structures considered, where $Cost = \begin{bmatrix} c_{1|1} & c_{1|2} & c_{1|3} \\ c_{2|1} & c_{2|2} & c_{2|3} \\ c_{3|1} & c_{3|2} & c_{3|3} \end{bmatrix}$, are $Cost_1 = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}$ and $Cost_2 = \begin{bmatrix} 0 & 2 & 5 \\ 1 & 0 & 3 \\ 1 & 3 & 0 \end{bmatrix}$. All prevalences remained the same ($p_j = 1/3$). Coverage

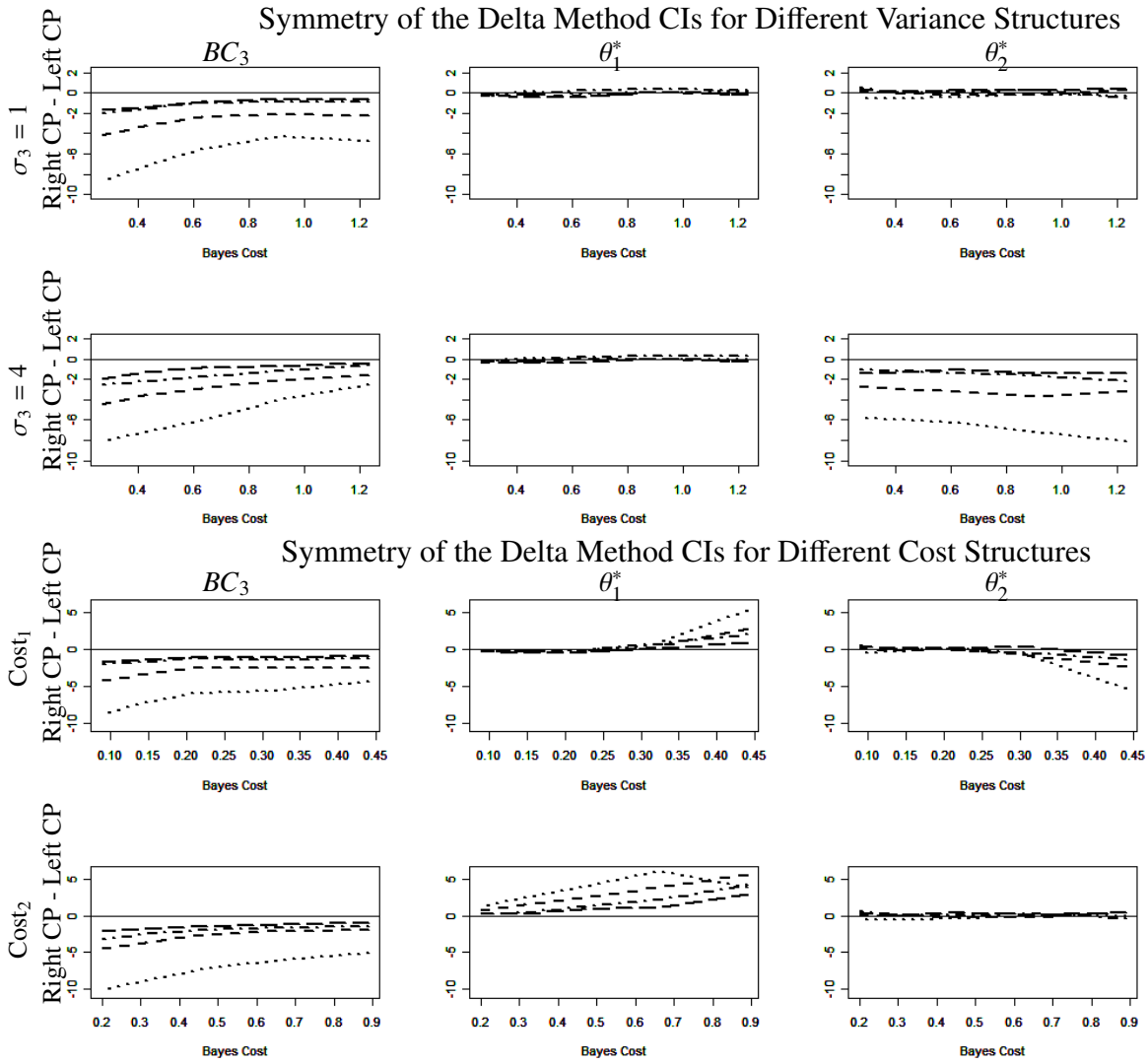

Figure 3.1: Plots of the difference between right and left coverage probability (CP) for the delta method CIs around BC_3, θ_1^* , and θ_2^* to consider the symmetry of the CIs for $n_j = 10$ (dotted line), $n_j = 50$ (dashed line), $n_j = 100$ (dash-dot line), and $n_j = 250$ (long dash line). Perfect symmetry would result in values of zero, and negative values indicated the right coverage is worse than the left coverage.

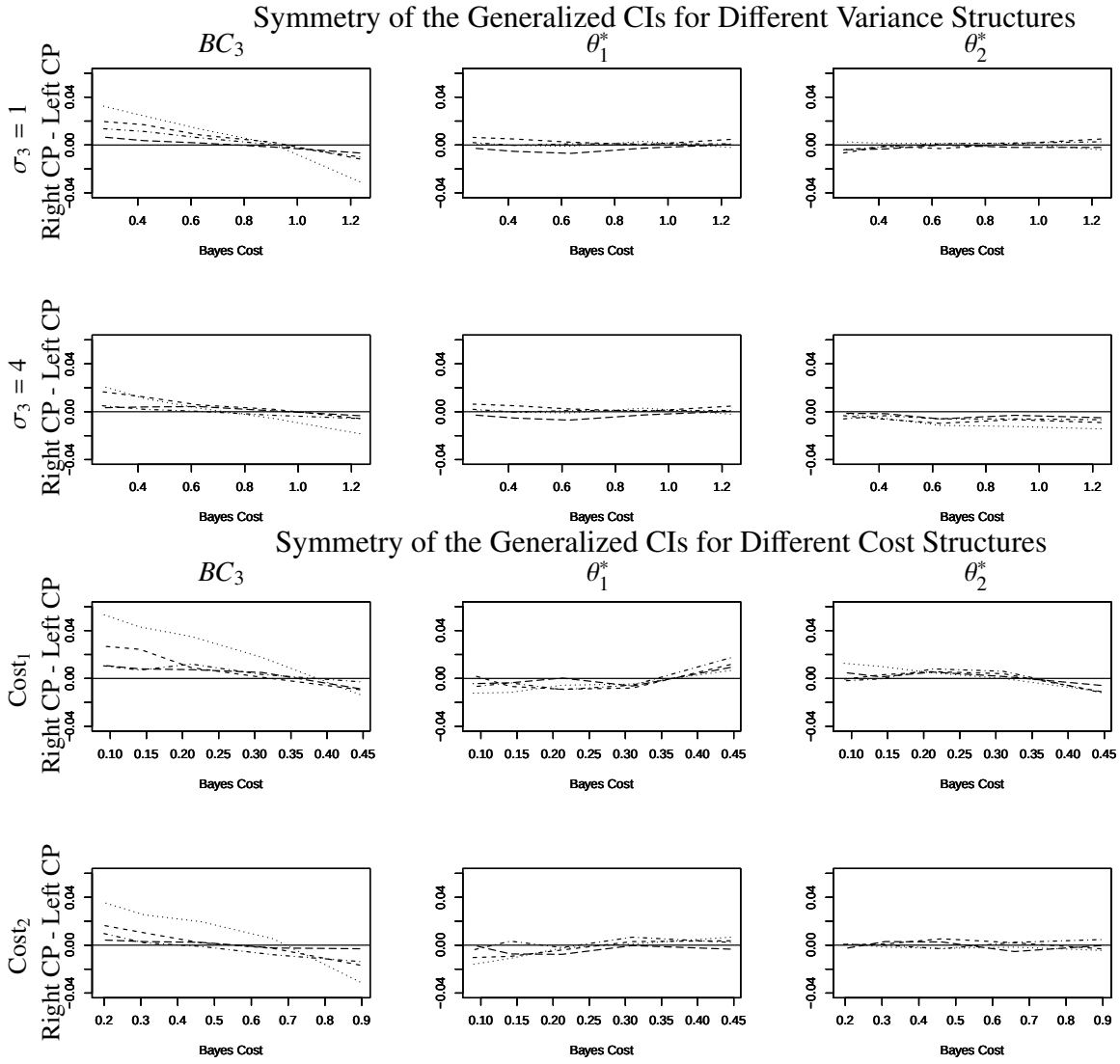

Figure 3.2: Plots of the difference between right and left coverage probability (CP) for the GCIs around BC_3, θ_1^* , and θ_2^* to consider the symmetry of the CIs for $n_j = 10$ (dotted line), $n_j = 50$ (dashed line), $n_j = 100$ (dash-dot line), and $n_j = 250$ (long dash line). Perfect symmetry would result in values of zero, and negative values indicated the right coverage is worse than the left coverage.

probabilities for the 95% delta, generalized and bootstrapped CIs for BC_3, θ_1^* , and θ_2^* are presented in Tables B.7 through B.9. The symmetry of the delta and generalized CIs are presented in the last two rows of Figures 3.1 and 3.2, respectively.

Similar to the different variance structures considered in Section 3.5.1, the varying cost structures do not have a noticeable impact on the bias or coverage probability for either BC or the optimal thresholds for the delta method or generalized CIs. The GCIs continue to perform better than the other methods with respect to coverage around BC at small n. For larger n, the delta and BCa CIs both perform well with respect to coverage. The CIs around the optimal thresholds have larger length at small n and larger BC_3 values. Once again, the GCIs are the only CIs achieving the desired coverage for $n_j = 10$. For $n_j \geq 50$, all methods perform well with respect to coverage and have similar lengths.

The symmetry of the delta method CIs around the optimal thresholds is altered by the different cost structures. $Cost_1$ distributes the costs evenly across class one and class three, resulting in asymmetry around both optimal thresholds (θ_1^* is skewed right and θ_2^* is skewed left). $Cost_2$ assigns the highest costs on class three, second highest costs on class two and lowest costs on class one. This results in the delta method CI around θ_1^* to become skewed right, while having no impact on the symmetry of θ_2^* . Asymmetry of the delta method CIs around the optimal threshold caused by varying the cost structure was also noted in [30] (for a two-class scenario, found using the GYI). Interestingly, the delta method CIs around BC maintain a fairly constant asymmetry for all cost and variance structures considered (Figure 3.1, column 1). Finally, although the varying cost structures have some impact on the symmetry of the GCIs (Figure 3.2, rows 3 and 4), this change is once again much smaller than that observed with the delta method CIs.

3.6 Summary

The delta method and generalized CIs were derived for BC under the assumption of a single feature used for classification that is independently and normally distributed for each class in a multi-state classification setting. Using simulations, the delta method CIs are shown to have good coverage for sample sizes of 50 or larger within each class and the GCIs are shown to have good coverage for sample sizes of 10 or more within each class, when the assumption of normality is

met for both methods. Notably, the BCa bootstrap CI with a parametric estimate of BC performs very similar to delta method CI around BC for most scenarios when the feature is normal. The performance of the delta method, generalized, and BCa bootstrapped CIs around BC is degraded when the assumption of normality is not met (for untransformed distributions). Performance of the derived CI methods around the optimal thresholds is also studied in the simulation. The delta method and generalized CIs around the optimal thresholds perform well when the assumption of normality is met, and are more robust to changes in variance than the three bootstrap methods considered. When the normality assumption is not met, all CI methods around the optimal thresholds have poor performance, with the performance being slightly better for specific normal mixture distributions. In addition, all CI methods are shown to be more robust to departures from normality for CIs around BC when compared to the same CI methods around the optimal thresholds. Finally, the GCIs performed the best with respect to coverage for a normally distributed feature (all sample sizes) with similar lengths as the other methods. The GCI have slightly longer lengths for the small sample size scenarios ($n_j = 10$). However, the GCIs are the only method achieving the desired coverage for this sample size, and therefore the longer length is expected. Therefore, the GCIs are recommended for all sample sizes and costs, and the delta method CIs may also be used for any large sample size and cost scenario (both for a normally distributed feature).

When all $c_{i|j}p_j$ are equal, for $i \neq j$, performance of CIs around BC may be compared to CI methods for J, as these two metrics measure performance equivalently (see Theorem 1). Currently, there are more CI methods available for J, although notably usually only for two classes. In general, the literature which proposes CIs for J use inconsistent bootstrap methods for comparison of the new methods' performance, making comparisons across all methods difficult. In [36], several estimates of J were considered for the bootstrap CIs (parametrically, empirically, Gaussian kernel smoothing, and kernel smoothing with Sheather-Jones algorithm), however, only BP CIs were presented which were shown in this chapter to only perform well for very large samples when considering a CI around BC. In [64], the empirical and parametric estimate of BC were both considered for the bootstrap CIs, however again, only the BP CI was utilized. Three bootstrap CI methods (BP, AN, and BCa) were used in [56], however, only empirical estimates of J and the optimal thresholds were

used with the bootstraps instead of parametric estimates for comparison to the delta method CIs (with the bootstrap CI performing worse than the delta method CI). This is expected, since in [45], it was shown that when classification systems result from a normally distributed feature, an empirical estimate of J has larger bias than the parametric estimate. Finally, in [33], a parametric resample is used with the assumption of a single feature with independent normal distributions for each class, with fairly good results. All other methods discussed utilize a nonparametric resample of the data. It would seem that if the feature is assumed to be normally distributed, then such an assumption should extend to the comparative methods, which suggests that the parametric estimation of BC with a BCa bootstrap is the appropriate bootstrap method. In this chapter it was shown that for a CI around BC (or similarly J) a parametric estimate of BC with a BCa bootstrap CI performs very similar to the delta method CI, and therefore is recommended for use when implementing a bootstrap CI for BC with a normally distributed feature. This bootstrap method outperforms those with empirical estimates of BC or J as the empirical estimate results in a higher bias compared to the parametric estimate [36, 45]. However, the BCa CI does not perform as well as the GCI around BC for a normally distributed feature with small sample sizes ($n_j = 10$) or as well as the delta method CI for accurate classification scenarios with a gamma distributed feature.

Another result of interest from the simulation study is the consistency of the delta method CI around BC to be skewed left (under all distributional and cost structures considered). This appears to be a result of the BC metric being the minimization of the misclassification rates (subject to prevalence and cost multipliers). This skewness is not seen with the GCI. Although asymmetrical, the delta method CIs still achieve the desired coverage probabilities and therefore the asymmetry is not necessarily a point of concern. The delta method CIs around the optimal thresholds are symmetric for equal variance of the feature's distribution for each class and balanced cost structures. Changing the variance or cost structure will impact the symmetry of the optimal thresholds' delta method CIs, as might be expected. In [30], asymmetry of the delta method CI around the optimal threshold was also noted when using the GYI for varying values of R (the prevalence and cost/benefit ratio) in the two-class framework with a normally (or log-normally) and independently distributed feature. Again, although the symmetry of the delta method CI is changed, the coverage still

meets desired levels for large n. Much smaller asymmetries are observed with the GCIs. Because symmetry is expected to behave similarly for other comparable scenarios of BC and threshold CIs, it will not be examined further in other methods.

Numerical estimation of the partial derivatives required for implementation of the delta method makes the application of the delta method CIs in this chapter (especially for $k > 3$ classes) more tangible. The methods presented in this chapter are especially useful since transformation techniques, such as the employed Box-Cox transformation, can be used to transform data to normality in order to meet the required assumptions so long as the underlying distributions lie in the Box-Cox family [45]. In Section 3.5, it is shown that the delta and generalized CIs around BC perform well and the CIs for the optimal thresholds do not perform well with respect to coverage for data transformed to normality. This further illustrates the usefulness of the CI around BC for choosing the best classifying feature, even when the optimal thresholds require further study to be determined accurately.

IV. Nonparamteric Confidence Intervals

4.1 Introduction

A CI for BC that does not require information about the structure of the classification system or feature distributions is derived in this chapter for any k classes. A nonparametric method for constructing a CI around BC is useful because small data sets or classifiers where distributional assumptions are not suitable occur regularly [5, 35, 37, 38, 57, 59]. Current nonparametric methods for J require large sample sizes (see Section 2.5.1). Although no distributional assumption is placed on the underlying feature(s), the classification outcomes from each class resulting from a fixed $\boldsymbol{\theta} \in \Theta$ are modeled with independent multinomial distributions.

This nonparametric CI around BC is developed in Section 4.2 using the fiducial argument. Available bootstrap methods that may be used in the nonparametric framework for constructing a CI around BC are presented in Section 4.3. In Section 4.4, simulations are used to demonstrate the performance of the newly developed method in Section 4.2 and compare its performance to other available CIs around BC in two- and three-class scenarios. Scenarios where the underlying classification system is unknown and scenarios with known normal distributions are considered. In Section 4.5, the newly developed method is compared further with available methods for developing simultaneous CIs around multinomial probabilities. Section 4.6 contains a summary of the results.

4.2 Fiducial Intervals

This section develops a CI for BC that requires no underlying distributional assumptions on the classification system. This CI is developed using the fiducial approach which was first introduced in 1930 by R.A. Fisher in his paper, "Inverse Probability" [21]. The fiducial argument has been used successfully for similar inference on statistical parameters [31, 74, 78], one very popular example being the Clopper-Pearson CI [3] for a binomial proportion (see Section 2.7.1.1) [13, 72]. The method developed in this section may be implemented for any (small) sample, k-class classification system and has a minimum coverage of $(1-\alpha)100\%$.

[3] Or fiducial interval, as these two terms are used interchangeably by Clopper and Pearson [13, 72]

The proposed CI requires only the observed classification outcomes, and assumes the outcomes are distributed multinomial. Section 4.2.1 derives the proposed method using the fiducial argument for the k-class BC with all $c_{i|j}p_j$ equal, for $i \neq j$. In Section 4.2.2, the method is extended for BC with unequal costs and prevalences. An algorithm for computing the upper and lower bounds is presented in Section 4.2.3 and an equivalence to a multiple of the Clopper-Pearson CI under specific conditions is also presented in Section 4.2.3.2. Finally, this method may be used equivalently for J, which is shown in Section 4.2.4.

Definition 3 (Fiducial Interval)**.** *A* $(1-\alpha)100\%$ *fiducial interval for a parameter* θ *is the set of values of* θ *which could have given rise to the observed value* Y=y *with the specified probability* $1-\alpha$*, and* $Y = t(X_1, ..., X_n)$ *a statistic from the random sample* $X_1, ..., X_n$ *with distribution* $F_Y(y|\theta)$ *[72].*

Therefore a $(1-\alpha)100\%$ fiducial interval for a parameter θ derived from an observed statistic $Y = t(X_1, ..., X_n)$ can be found as the solutions for θ_L and θ_U in the following equations [72]:

$$Pr(Y \geq y \mid \theta_L) = \frac{\alpha}{2} \tag{4.1}$$

$$Pr(Y \leq y \mid \theta_U) = \frac{\alpha}{2} \tag{4.2}$$

4.2.1 Bayes Cost with Equal Weights.

Initially, it is assumed that all $c_{i|j}p_j$ are equal to one, for $i \neq j$[4]. Then BC can be expressed as the sum of the $k^2 - k$ misclassification probabilities resulting from the k-class classification system. Here, the minimization is excluded because it is assumed the classifier is applied at its optimal setting, or more generally at a fixed setting. Specifically,

$$BC = \sum_{\substack{i=1 \\ i \neq j}}^{k} \sum_{j=1}^{k} p_{i|j} \tag{4.3}$$

where each $p_{i|j}$ is the probability of classifying an observation from class j as class i ($j = 1, \ldots, k$ and $i = 1, \ldots, k$) . The statistic used to estimate BC is $Y = \widehat{BC}$ where

$$Y = \sum_{\substack{i=1 \\ i \neq j}}^{k} \sum_{j=1}^{k} \frac{X_{i|j}}{n_j} \tag{4.4}$$

[4] As long as all multipliers on each misclassification probability are equal, the multiplier can be scaled to one without changing the classification outcomes.

Each $X_{i|j}$ is a multinomial random variable representing the number of observations classified as the i^{th} class when their true class is j , and n_j is the total number of observations for the j^{th} class.

The statistic Y is a function of discrete random variables representing a projection into the one dimensional rational space ($\mathbb{Q}$), and can be ordered (possibly with ties). From Equations 4.1 and 4.2, the $(1-\alpha)100\%$ fiducial interval for BC from an observed statistic $Y = y$ is determined by the values of BC_L and BC_U that are the solutions to the following equations:

$$Pr(Y \geq y \mid BC_L) = \frac{\alpha}{2} \tag{4.5}$$

$$Pr(Y \leq y \mid BC_U) = \frac{\alpha}{2} \tag{4.6}$$

To find these solutions, the probability distribution of Y with respect to BC must be determined. For each class, $\mathbf{X_j} = (X_{1|j}, \ldots, X_{k|j}) \sim multinomial(\mathbf{p_j}, n_j)$, where each $X_{i|j}$ is a nonnegative integer and $\sum_{i=1}^{k} X_{i|j} = n_j$. The multinomial pmf for $\mathbf{X_j}$ is of the form

$$f_{\mathbf{X_j}}(\mathbf{x_j}) = n_j! \prod_{i=1}^{k} \frac{p_{i|j}^{x_{i|j}}}{x_{i|j}!} \tag{4.7}$$

[12]. Therefore, the joint pmf for all k^2 random variables, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_k)$, from the k independent multinomial distributions resulting from the k-class classifier is

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{p}) &= \prod_{j=1}^{k} f_{\mathbf{X_j}}(\mathbf{x_j}) \\ &= \prod_{i=1}^{k} \prod_{j=1}^{k} n_j! \frac{p_{i|j}^{x_{i|j}}}{x_{i|j}!} \end{aligned} \tag{4.8}$$

Let $\mathcal{S}$ represent the probability parameter space for the entire experiment, $\mathbf{p} \in \mathcal{S} = \{\mathbf{p} = (\mathbf{p_1}, \ldots, \mathbf{p_k}) : \mathbf{p_j} = (p_{1|j}, \ldots, p_{k|j}), p_{i|j} \geq 0$, and $\sum_{i=1}^{k} p_{i|j} = 1\}$. Also let $\mathcal{A}$ be the joint multinomial sample space that is the set of $1 \times k^2$ sized vectors where $\mathcal{A} = \{\mathbf{x} = (\mathbf{x_1}, \ldots, \mathbf{x_k}) : \mathbf{x_k} = (x_{1|j}, \ldots, x_{k|j}), x_{i|j} \in \mathbb{Z}^+, \sum_{i=1}^{k} x_{i|j} = n_j\}$. For a single multinomial distribution, there are

$$\begin{pmatrix} n+k-1 \\ n \end{pmatrix} \tag{4.9}$$

distinct elements in the sample space. For example, for $k = 3$ outcomes and $n = 2$ observations from a single multinomial experiment, there are six elements (shown in Table 4.1). In addition, for

Table 4.1: *The multinomial sample space for 3 outcomes and* $n = 2$ *. Each row represents a potential draw from the multinomial experiment.*

$X_{1\|j}$	$X_{2\|j}$	$X_{3\|j}$
2	0	0
1	0	1
1	1	0
0	0	2
0	1	1
0	2	0

a k-class classification system, there are

$$\prod_{j=1}^{k} \begin{pmatrix} n_j + k - 1 \\ n_j \end{pmatrix} \tag{4.10}$$

distinct ways of sampling from this joint multinomial experiment (ie. number of elements in $\mathcal{A}$). Clearly, as k and each n_j increase, this sample space becomes large. For the previous example where $k = 3$, if each $n_j = 2$ there are 216 distinct ways of sampling from the joint multinomial experiment.

With the assumption of all $c_{i|j}p_j$ being equal, for $i \neq j$, the sum of the $k - 1$ misclassification rates for each class may be treated as a total misclassification rate for that class. BC can then be defined using the total misclassifications only, as it is unnecessary to distinguish between the types of misclassifications (e.g. $X_{2|1}$ vs $X_{3|1}$). For simplicity of notation, the sum of the $k - 1$ misclassification probabilities from each class is denoted $p_{j^c|j}$:

$$p_{j^c|j} = \sum_{\substack{i=1 \\ i \neq j}}^{k} p_{i|j} \tag{4.11}$$

The total number of misclassified observations from each class is denoted $X_{j^c|j}$:

$$X_{j^c|j} = \sum_{\substack{i=1 \\ i \neq j}}^{k} X_{i|j} \tag{4.12}$$

The independent multinomial distributions can be collapsed into k independent binomial distributions, with the total misclassifications representing success and the correct classifications

representing failure in each class. Thus for each fixed j ,

$$\begin{aligned} \textstyle\sum_{i=1}^{k} X_{i|j} &= n_j \\ &\Rightarrow X_{i=j|j} + \textstyle\sum_{i\neq j}^{k} X_{i|j} = n_j \\ &\Rightarrow X_{j^c|j} = n_j - X_{i=j|j} \\ &\Rightarrow \#\text{ of misclassifications} = n_j - \#\text{ of correct classifications} \\ &\Rightarrow X_{j^c|j} \sim \text{Bin}(n_j, p_{j^c|j}) \end{aligned} \tag{4.13}$$

Considering only the total misclassifications for each class (modeled as independent binomial random variables), the size of the sample space for the classification system is reduced to

$$\prod_{j=1}^{k}(n_j + 1) \tag{4.14}$$

The reduction of the sample space is demonstrated in Table 4.2 with a single class from the previous example, where $k = 3$ and $n = 2$. In this example, the number of elements in the sample space for one class is reduced from six (multinomial sample space) to three (binomial sample space).

Therefore, the joint pmf for the k^2 independent multinomial random variables, $\mathbf{X} = (X_{1|1}, X_{2|1}, \ldots, X_{k-1|k}, X_{k|k})$, can be expressed using the joint pmf for k independent binomial random

Table 4.2: *A multinomial sample space reduced to a binomial sample space for 3 outcomes and* $n = 2$ *. Each row represents a potential draw from the experiment (assuming the truth class is 1, therefore* $X_{1|1}$ *is the correct classification)*

$X_{1\|1}$	$X_{2\|1}$	$X_{3\|1}$		$X_{1\|1}$	$X_{2\|1} + X_{3\|1}$		$X_{1\|1}$	$X_{j^c\|1}$
2	0	0		2	$0+0=0$		2	$0 = n-2$
1	0	1		1	$0+1=1$		1	$1 = n-1$
1	1	0	$\rightarrow$	1	$1+0=1$	$\rightarrow$	0	$2 = n-0$
0	0	2		0	$0+2=2$			
0	1	1		0	$1+1=2$			
0	2	0		0	$2+0=2$			

variables, $\mathbf{X} = (X_{1^c|1}, \ldots, X_{k^c|k})$, where each $X_{j^c|j}$ is a nonnegative integer and $0 \leq X_{j^c|j} \leq n_j$:

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{p}) &= \prod_{j=1}^{k} f_{\mathbf{X_j}}(\mathbf{x_j}) \\ &= \prod_{j=1}^{k} n_j! \frac{\left(p_{j^c|j}\right)^{x_{j^c|j}}}{x_{j^c|j}!} \frac{\left(1 - p_{j^c|j}\right)^{x_{j|j}}}{x_{j|j}!} \\ &= \prod_{j=1}^{k} \binom{n_j}{x_{j^c|j}} \left(p_{j^c|j}\right)^{x_{j^c|j}} \left(q_{j^c|j}\right)^{\left(n_j - x_{j^c|j}\right)} \end{aligned} \tag{4.15}$$

Here, $q_{j^c|j} = (1 - p_{j^c|j})$ and $\mathbf{p} = (p_{1^c|1}, \ldots, p_{k^c|k})$ is a vector of the k total misclassification probabilities from the classification system.

Recall $\mathcal{A}$ is the joint multinomial sample space. Let the reduced sample space, $\mathcal{B}$, be the joint binomial sample space that is the set of $1 \times k$ sized vectors where $\mathcal{B} = \{\mathbf{x} = (x_{1^c|1}, \ldots, x_{k^c|k}) : x_{j^c|j} \in \mathbb{Z}^+, x_{j^c|j} \leq n_j\}$. Then the sample space for $Y = \widehat{BC}$ is $\mathcal{Y} = \{y : y = \sum_{i=1, i\neq j}^{k} \sum_{j=1}^{k} \frac{x_{i|j}}{n_j}, \mathbf{x} \in \mathcal{B}\}$. Therefore, the pmf of Y with respect to the binomial probabilities $\mathbf{p} = (p_{1^c|1}, \ldots, p_{k^c|k})$ can be written in terms of the joint binomial distribution as

$$\begin{aligned} f_Y(y \mid \mathbf{p}) &= P(Y = y \mid \mathbf{p}) \\ &= P\left(\sum_{\substack{i=1 \\ i\neq j}}^{k} \sum_{j=1}^{k} \frac{X_{i|j}}{n_j} = y \mid \mathbf{p} \right) \\ &= \sum_{\substack{\mathbf{x}\in\mathcal{B} \\ Y=y}} f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{p}) \end{aligned} \tag{4.16}$$

where $f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{p})$ is defined in Equation 4.15. The last line in Equation 4.16 is a summation because it is possible to have more than one $\mathbf{x} \in \mathcal{B}$ that results in $Y = y$ (these are ties in the ordered sample space). For example, if $k = 3$ and each $n_j = 2$, an observed $\widehat{BC} = 0.5$ will occur if there is one misclassification out of the total six observations. There are three ways of observing only one misclassification from this experiment, resulting in the ties in the sample space for $Y = y$. These ties are shown in Table 4.3.

Using Equation 4.16, the CDF of Y with respect to $\mathbf{p} = (p_{1^c|1}, \ldots, p_{k^c|k})$ is

$$F_Y(y \mid \mathbf{p}) = \sum_{t=0}^{y} f_Y(t \mid \mathbf{p}) = \sum_{t=0}^{y} \sum_{\substack{\mathbf{x}\in\mathcal{B} \\ Y=t}} f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{p}) \tag{4.17}$$

Table 4.3: *Ties in the joint binomial sample space for 3 classes,* $n_j = 2$ *, and* $\widehat{BC} = 0.5$ *. Each row represents an element from the joint binomial experiments.*

Class 1 ($X_{1^c\|1}$)	Class 2 ($X_{2^c\|2}$)	Class 3 ($X_{3^c\|3}$)	$\widehat{BC}$
0	0	1	0.5
0	1	0	0.5
1	0	0	0.5

For each fixed BC, there exists infinite $\mathbf{p} = (p_{1^c|1}, \ldots, p_{k^c|k})$ such that $\mathbf{p}^T\mathbf{1} = BC$ (where $\mathbf{1}$ is a $k \times 1$ sized vector of ones), resulting in different values of $F_Y(y \mid \mathbf{p})$ for a given BC and observed $y = \widehat{BC}$ (except for the trivial cases where $BC = 0$ or $BC = k$). This makes finding a unique solution for the fiducial bounds on BC, given in Equations 4.5 and 4.6, impossible. To demonstrate multiple values of $F_Y(y \mid \mathbf{p})$ for each fixed BC, an example where $y = 0.5$ (left) and $y = 1$ (right) is shown in Figure 4.1. This example plots $F_Y(y \mid \mathbf{p})$ (Equation 4.17, plotted with black dots) against BC with multiple $\mathbf{p}$ ($\mathbf{p}^T\mathbf{1} = BC$). Therefore, define $F_Y^1(y \mid BC)$ to be the maximum value of $F_Y(y \mid \mathbf{p})$ for each fixed $BC = \mathbf{p}^T\mathbf{1}$ and $F_Y^2(y \mid BC)$ to be the minimum value of $F_Y(y \mid \mathbf{p})$ for each fixed $BC = \mathbf{p}^T\mathbf{1}$. Then these two functions are one-to-one and onto from BC to the $F_Y(y \mid \mathbf{p})$ space, and unique solutions for the fiducial bounds can be found. These two new functions are shown in Figure 4.1 where the blue line is $F_Y^1(y \mid BC)$ and the red line is $F_Y^2(y \mid BC)$. These functions can be expressed using Equation 4.17 as

$$F_Y^1(y \mid BC) = \max_{\substack{\mathbf{p}:\mathbf{p}^T\mathbf{1}=BC \\ BC\in\mathcal{BC}}} \left[\sum_{t\le y} \sum_{\substack{\mathbf{x}\in\mathcal{B} \\ Y=t}} f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{p}) \right] \tag{4.18}$$

$$F_Y^2(y \mid BC) = \min_{\substack{\mathbf{p}:\mathbf{p}^T\mathbf{1}=BC \\ BC\in\mathcal{BC}}} \left[\sum_{t\le y} \sum_{\substack{\mathbf{x}\in\mathcal{B} \\ Y=t}} f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{p}) \right] \tag{4.19}$$

where $\mathcal{BC}$ is the BC sample space such that $\mathcal{BC} = \{BC : BC = \mathbf{p}^T\mathbf{1}, \mathbf{p} = (p_{1^c|1}, \ldots, p_{k^c|k}), p_{i^c|j} \in [0, 1]\}$.

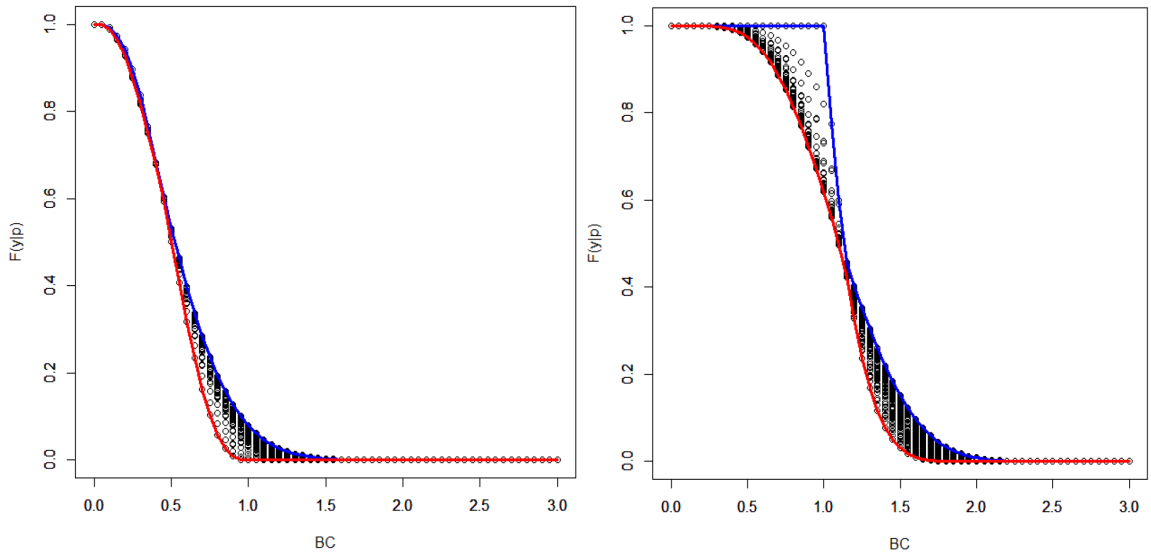

Figure 4.1: Example of $F_y(y \mid \mathbf{p})$ vs BC for an observed $y = 0.5$ (left) and $y = 1$ (right).

Combining Equations 4.18 and 4.19 with Equations 4.5 and 4.6, the lower (BC_L) and upper (BC_U) bounds for the $(1-\alpha)100\%$ fiducial interval for BC from an observed statistic y are:

$$BC_L = \sup\left\{BC \in \mathcal{BC} \text{ such that } 1 - \min_{\mathbf{p}:\mathbf{p}^T\mathbf{1}=BC}\left[\sum_{t\leq y^*}\sum_{\substack{\mathbf{x}\in\mathcal{B}\\ y=t}} f_{\mathbf{X}}(\mathbf{x}\mid\mathbf{p})\right] \leq \frac{\alpha}{2}\right\} \tag{4.20}$$

$$BC_U = \inf\left\{BC \in \mathcal{BC} \text{ such that } \max_{\mathbf{p}:\mathbf{p}^T\mathbf{1}=BC}\left[\sum_{t\leq y}\sum_{\substack{\mathbf{x}\in\mathcal{B}\\ y=t}} f_{\mathbf{X}}(\mathbf{x}\mid\mathbf{p})\right] \leq \frac{\alpha}{2}\right\} \tag{4.21}$$

where y^* is the ordered value of $Y \in \mathcal{Y}$ directly less than y. When $y = 0$ or $y = k$, the lower bound is $BC_L = 0$ and the upper bound is $BC_U = k$, respectively. This is due to the fact that $Y \in [0, k]$ when all $c_{i|j}p_j$ are assumed equal to one, for $i \neq j$, making $Pr(Y \geq 0 \mid BC) = 1$ and $Pr(Y \leq k \mid BC) = 1$. The upper and lower bounds expressed in Equations 4.20 and 4.21 may be found by searching all $\mathbf{p}$ within a certain tolerance, which motivates using inequalities to meet the minimum coverage desired. The coverage of this CI is addressed in the following theorem.

Theorem 5. *The upper and lower bounds for BC given by*

$$BC_L = \sup\left\{BC \in \mathcal{BC} \text{ such that } 1 - \min_{\mathbf{p}:\mathbf{p}^T\mathbf{1}=BC}\left[\sum_{t\le y^*}\sum_{\substack{\mathbf{x}\in\mathcal{B}\\ y=t}} f_{\mathbf{X}}(\mathbf{x}\mid\mathbf{p})\right] \le \frac{\alpha}{2}\right\} \tag{4.20}$$

$$BC_U = \inf\left\{BC \in \mathcal{BC} \text{ such that } \max_{\mathbf{p}:\mathbf{p}^T\mathbf{1}=BC}\left[\sum_{t\le y}\sum_{\substack{\mathbf{x}\in\mathcal{B}\\ y=t}} f_{\mathbf{X}}(\mathbf{x}\mid\mathbf{p})\right] \le \frac{\alpha}{2}\right\} \tag{4.21}$$

create a $(1-\alpha)100\%$ *fiducial interval around BC when weights on misclassification costs are equal with a confidence coefficient of at least* $(1-\alpha)100\%$.

Proof. *Let* $BC \in \mathcal{BC}$, $y = \widehat{BC}$, *and* $\mathbf{p} = (p_{1^c|1}, \ldots, p_{k^c|k})$ *be k joint binomial total misclassification probabilities from a k-class classification system. Since* $BC = \sum_{i\neq 1}^{k}\sum_{j=1}^{k} p_{i|j}$, *any small increase of* ϵ *in any one* $p_{i|j}$ *will result in an increase of* ϵ *in BC. For the upper bound this results in,*

$$Pr(Y \le y \mid BC_U) \le \max_{\substack{\mathbf{p}:\mathbf{p}^T\mathbf{1}=BC_U\\ BC_U\in\mathcal{BC}}}\left[\sum_{t\le y}\sum_{\substack{\mathbf{x}\in\mathcal{B}\\ Y=t}} f_{\mathbf{X}}(\mathbf{x}\mid\mathbf{p})\right]$$

$$\Rightarrow Pr(Y \le y \mid BC_U) \le \frac{\alpha}{2} \tag{4.22}$$

Now let y^* *be the ordered value of* $Y \in \mathcal{Y}$ *directly less than y. Then for the lower bound,*

$$Pr(Y \ge y \mid BC_L) \le 1 - \min_{\substack{\mathbf{p}:\mathbf{p}^T\mathbf{1}=BC_L\\ BC_L\in\mathcal{BC}}}\left[\sum_{t\le y^*}\sum_{\substack{\mathbf{x}\in\mathcal{B}\\ Y=t}} f_{\mathbf{X}}(\mathbf{x}\mid\mathbf{p})\right]$$

$$\Rightarrow Pr(Y \ge y \mid BC_L) \le \frac{\alpha}{2} \tag{4.23}$$

The confidence coefficient for any CI is given generally in [72] as

$$Pr(\theta_L < \theta < \theta_U) = Pr(Y < y \mid \theta_L) - Pr(Y \le y \mid \theta_U) \tag{4.24}$$

Therefore for the fiducial interval around BC

$$\begin{aligned} Pr(BC_L < BC < BC_U) &= Pr(Y < y \mid BC_L) - Pr(Y \le y \mid BC_U) \\ &= 1 - Pr(Y \ge y \mid BC_L) - Pr(Y \le y \mid BC_U) \\ &\ge 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha \\ &\Rightarrow [Pr(BC \in [BC_L, BC_U] \mid y)] \ge 1 - \alpha \end{aligned} \tag{4.25}$$

□

The proof for Theorem 5 does not depend on the sample size used to develop the fiducial interval. Therefore, the minimum desired coverage of $(1-\alpha)100\%$ will be met for any sample size, making this method appropriate for small samples where approximate methods fail to achieve the necessary coverage. Also, this method relies on ordering the sample space of the k joint independent binomial distributions. This sample space becomes large as k and each n_j increase, making this method, in addition to being suitable, more practical for small samples.

4.2.2 Bayes Cost with Unequal Weights.

When all $c_{i|j}p_j$ are not equal, for $i \neq j$, the method for finding the fiducial interval around BC becomes more involved compared to when all multipliers are equal. First, the outcomes from the classification system can no longer be reduced to binomial random variables. BC is more generally defined in this scenario as,

$$BC = \sum_{\substack{i=1 \\ i\neq j}}^{k} \sum_{j=1}^{k} c_{i|j} p_j p_{i|j} \tag{4.26}$$

where each $p_{i|j}$ is the probability of classifying an observation from class j as class i , p_j is the prevalence of class j , and $c_{i|j}$ is the cost associated with classifying class j as class i ($j = 1, \ldots, k$ and $i = 1, \ldots, k$), and the minimization is excluded because it is assumed the classification system is applied at its optimal settings. The statistic used to estimate BC is $Y = \widehat{BC}$,

$$Y = \sum_{\substack{i=1 \\ i\neq j}}^{k} \sum_{j=1}^{k} c_{i|j} p_j \frac{X_{i|j}}{n_j} \tag{4.27}$$

Because each misclassification with respect to truth must be considered uniquely (for example, $X_{2|1}$ vs $X_{3|1}$), the k^2 random variables $\mathbf{X} = (X_{1|1}, X_{2|1}, \ldots, X_{k-1|k}, X_{k|k})$ must be modeled with the multinomial distribution[5]

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{p}) &= \prod_{j=1}^{k} f_{\mathbf{X_j}}(\mathbf{x_j}) \\ &= \prod_{i=1}^{k} \prod_{j=1}^{k} n_j! \frac{p_{i|j}^{x_{i|j}}}{x_{i|j}!} \end{aligned} \tag{4.8}$$

[5]To reduce computation time when searching for the lower and upper CI bounds on BC, if any of the k classes have equal weights on the class misclassifications, this class's total misclassification may be modeled as binomial, and the binomial pmf may be used for that specific $f_{\mathbf{X_j}}(\mathbf{x_j})$.

Here $\mathbf{X} = (X_{1|1}, X_{2|1}, \ldots, X_{k-1|k}, X_{k|k})$ and $\mathbf{p} \in \mathcal{S} = \{\mathbf{p} = (\mathbf{p_1}, \ldots, \mathbf{p_k}) : \mathbf{p_k} = (p_{1|j}, \ldots, p_{k|j}), p_{i|j} \geq 0$, and $\sum_{i=1}^{k} p_{i|j} = 1\}$. Again let $\mathcal{A}$ be the joint multinomial sample space defined in Section 4.2.1. Similar to the method in Section 4.2.1, the CDF of Y with respect to the multinomial misclassification probabilities can be written as

$$F_Y(y \mid \mathbf{p}) = \sum_{t=0}^{y} f_Y(t \mid \mathbf{p}) = \sum_{t=0}^{y} \sum_{\substack{\mathbf{x}\in\mathcal{A} \\ Y=t}} f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{p}) \tag{4.28}$$

where $f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{p})$ is defined in Equation 4.8 and $\mathcal{Y} = \{y : y = \sum_{i=1, i\neq j}^{k} \sum_{j=1}^{k} c_{i|j} p_j \frac{x_{i|j}}{n_j}, \mathbf{x} \in \mathcal{A}\}$.

When all $c_{i|j}p_j$ are not equal, for $i \neq j$, BC is no longer defined simply as the sum of the misclassification probabilities. Therefore, any small increase of ϵ in any one $p_{i|j}$ will not necessarily result in an increase of ϵ in BC. It is clear that when the weights are different, a small increase in any one $p_{i|j}$ will have a different impact on BC depending on the specific misclassification probability's cost and prevalence. Therefore if $F_Y^1(y \mid BC)$ and $F_Y^2(y \mid BC)$ are defined as they were for equal weights in Equations 4.18 and 4.19 in Section 4.2.1, the coverage probability of the CI will not be guaranteed for unequal costs of misclassification. Instead, a small adjustment is made to these definitions to ensure coverage for CI around BC with unequal costs or prevalences meets the desired level of $1 - \alpha$. Define two step functions,

$$F_Y^3(y \mid BC_U) = \max_{BC \geq BC_U} \left[F_Y^1(y \mid BC)\right] = \max_{BC \geq BC_U} \left\{ \max_{\substack{\mathbf{p}:\mathbf{p}^T\mathbf{c}=BC \\ BC\in\mathcal{BC}}} \left[\sum_{t\leq y} \sum_{\substack{\mathbf{x}\in\mathcal{A} \\ y=t}} f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{p}) \right] \right\} \tag{4.29}$$

$$F_Y^4(y \mid BC_L) = \max_{BC \leq BC_L} \left[1 - F_Y^2(y \mid BC)\right] = \max_{BC \leq BC_L} \left\{ 1 - \min_{\substack{\mathbf{p}:\mathbf{p}^T\mathbf{c}=BC \\ BC\in\mathcal{BC}}} \left[\sum_{t\leq y} \sum_{\substack{\mathbf{x}\in\mathcal{A} \\ y=t}} f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{p}) \right] \right\} \tag{4.30}$$

where $\mathbf{c}$ is a vector of the constant multipliers to be placed on each misclassification probability ($\mathbf{c} = (\mathbf{c_1}, \ldots, \mathbf{c_k})$, where $\mathbf{c_j} = (c_{1|j}p_j, \ldots, c_{k|j}p_j)$, and $c_{i|j}p_j \in \mathbb{R}^+$). A plot of $F_Y^1(y \mid BC)$, $1 - F_Y^2(y \mid BC)$, $F_Y^3(y \mid BC)$, and $F_Y^4(y \mid BC)$ is presented in Figure 4.2 for an example scenario where $\widehat{BC} = 0.99$ when $n_1 = 3$, $n_2 = 5$, $n_3 = 6$, and $Cost = \begin{bmatrix} 0 & 3 & 5 \\ 3 & 0 & 1 \\ 1 & 5 & 0 \end{bmatrix}$ (all p_j are assumed equal to $\frac{1}{3}$).

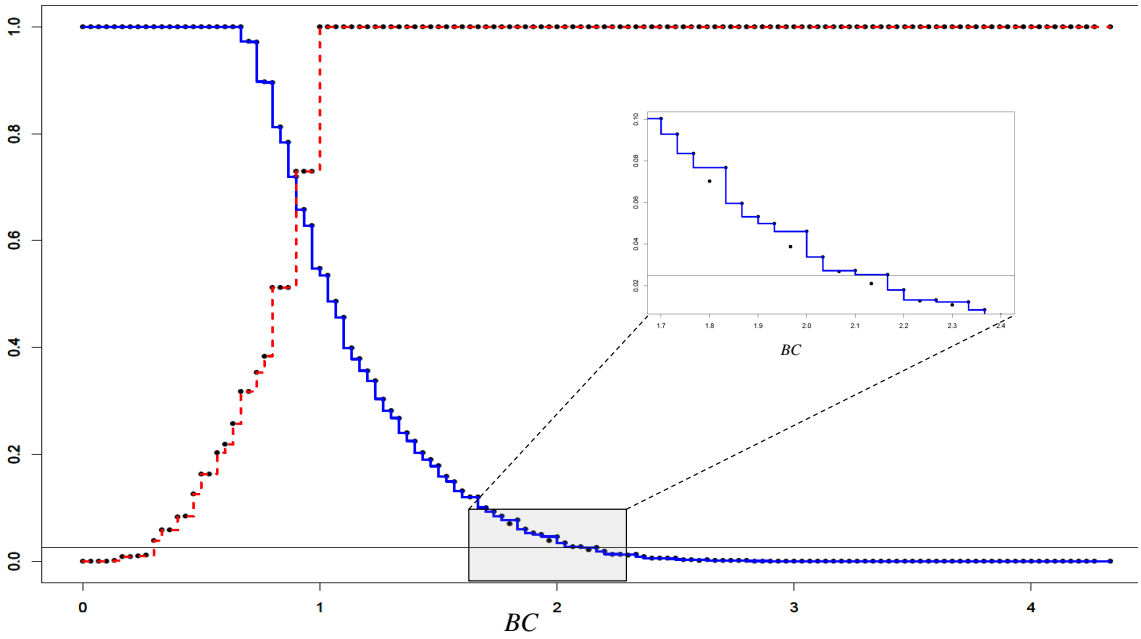

Figure 4.2: Example of $F_Y^3(y \mid BC_U)$ and $F_Y^4(y \mid BC_L)$ plotted vs BC for an observed $\widehat{BC} = 0.98889$ when $n_1 = 3$, $n_2 = 5$, $n_3 = 6$, and $Cost = \begin{bmatrix} 0 & 3 & 5 \\ 3 & 0 & 1 \\ 1 & 5 & 0 \end{bmatrix}$. The values for $F_Y^1(y \mid BC)$ and $1 - F_Y^2(y \mid BC)$ are plotted with the decreasing and increasing black dots, respectively. Then the values for $F_Y^3(y \mid BC_U)$ are plotted with the blue solid line and for $F_Y^4(y \mid BC_L)$ with the red dashed line. The black horizontal line is drawn at $\frac{\alpha}{2} = 0.025$.

The $(1-\alpha)100\%$ fiducial interval for BC from an observed statistic y is the BC_L and BC_U given by:

$$BC_L = \sup \left\{ BC \in \mathcal{BC} \text{ such that } \max_{\mathbf{p}:\mathbf{p}^T\mathbf{c} \le BC} \left[\sum_{t \ge y} \sum_{\substack{\mathbf{x} \in \mathcal{A} \\ y=t}} f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{p}) \right] \le \frac{\alpha}{2} \right\} \tag{4.31}$$

$$BC_U = \inf \left\{ BC \in \mathcal{BC} \text{ such that } \max_{\mathbf{p}:\mathbf{p}^T\mathbf{c} \ge BC} \left[\sum_{t \le y} \sum_{\substack{\mathbf{x} \in \mathcal{A} \\ y=t}} f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{p}) \right] \le \frac{\alpha}{2} \right\} \tag{4.32}$$

and $\mathcal{BC}$ is the parameter space for BC with unequal weights where $\mathcal{BC} = \{BC : BC = \mathbf{p}^T\mathbf{c}, \mathbf{c} = (\mathbf{c_1}, \ldots, \mathbf{c_k}), \mathbf{c_j} = (c_{1|j}p_j, \ldots, c_{k|j}p_j),$ and $c_{i|j}p_j \in \mathbb{R}^+, \mathbf{p} \in \mathcal{S}\}$. When $y = 0$ or $y = \sup\{\mathcal{Y}\}$, the lower bound is $BC_L = 0$ and the upper bound is $BC_U = \sup\{\mathcal{BC}\}$, respectively. This is due to the fact that $Y \in [0, \sup\{\mathcal{Y}\}]$ when all $c_{i|j}p_j$ are not equal, for $i \ne j$, and all $c_{i|j}p_j$ are greater than or equal

to zero, making $Pr(Y \geq 0 \mid BC) = 1$ and $Pr(Y \leq \sup\{\mathcal{Y}\} \mid BC) = 1$. The lower and upper bounds given in Equations 4.31 and 4.32 may be found by searching all $\mathbf{p}$ within a certain tolerance, which is why they are solved using inequalities to meet the desired minimum coverage. The coverage of this CI is addressed in the following theorem and proof.

Theorem 6. *The upper and lower bounds for BC given by*

$$BC_L = \sup\left\{BC \in \mathcal{BC} \text{ such that } \max_{\mathbf{p}:\mathbf{p}^T\mathbf{c}\leq BC}\left[\sum_{t\geq y}\sum_{\substack{\mathbf{x}\in\mathcal{A}\\ y=t}} f_{\mathbf{X}}(\mathbf{x}\mid\mathbf{p})\right] \leq \frac{\alpha}{2}\right\} \tag{4.31}$$

$$BC_U = \inf\left\{BC \in \mathcal{BC} \text{ such that } \max_{\mathbf{p}:\mathbf{p}^T\mathbf{c}\geq BC}\left[\sum_{t\leq y}\sum_{\substack{\mathbf{x}\in\mathcal{A}\\ y=t}} f_{\mathbf{X}}(\mathbf{x}\mid\mathbf{p})\right] \leq \frac{\alpha}{2}\right\} \tag{4.32}$$

create a $(1-\alpha)100\%$ *fiducial interval around BC when weights on misclassification rates are not equal with a confidence coefficient of at least* $(1-\alpha)100\%$.

Proof. *Let* $BC \in \mathcal{BC}$, $y = \widehat{BC}$, $\mathbf{p} \in \mathcal{S}$ *be the k joint multinomial probabilities from a k-class classification system, and* $\mathbf{c} = (\mathbf{c_1}, \ldots, \mathbf{c_k}), \mathbf{c_j} = (c_{1|j}p_j, \ldots, c_{k|j}p_j)$, *and* $c_{i|j}p_j \in \mathbb{R}^+$. *Also,* $BC = \sum_{i=1, i\neq j}^{k}\sum_{j=1}^{k} c_{i|j}p_j p_{i|j}$. *For the upper bound this results in,*

$$Pr(Y \leq y \mid BC_U) \leq \max_{BC\geq BC_U}\left\{\max_{\substack{\mathbf{p}:\mathbf{p}^T\mathbf{c}=BC\\ BC\in\mathcal{BC}}}\left[\sum_{t\leq y}\sum_{\substack{\mathbf{x}\in\mathcal{A}\\ y=t}} f_{\mathbf{X}}(\mathbf{x}\mid\mathbf{p})\right]\right\}$$

$$\Rightarrow Pr(Y \leq y \mid BC_U) \leq \frac{\alpha}{2} \tag{4.33}$$

Now let y^* *be the ordered value of* $Y \in \mathcal{Y}$ *directly less than y. Then for the lower bound,*

$$Pr(Y \geq y \mid BC_L) \leq \max_{BC\leq BC_L}\left\{1 - \min_{\substack{\mathbf{p}:\mathbf{p}^T\mathbf{c}=BC\\ BC\in\mathcal{BC}}}\left[\sum_{t\leq y^*}\sum_{\substack{\mathbf{x}\in\mathcal{A}\\ y=t}} f_{\mathbf{X}}(\mathbf{x}\mid\mathbf{p})\right]\right\}$$

$$\Rightarrow Pr(Y \geq y \mid BC_L) \leq \frac{\alpha}{2} \tag{4.34}$$

The confidence coefficient for the fiducial interval around any BC is

$$\begin{aligned} Pr(BC_L < BC < BC_U) &= Pr(Y < y \mid BC_L) - Pr(Y \leq y \mid BC_U) \\ &= 1 - Pr(Y \geq y \mid BC_L) - Pr(Y \leq y \mid BC_U) \\ &\geq 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha \\ &\Rightarrow [Pr(BC \in [BC_L, BC_U] \mid y)] \geq 1 - \alpha \end{aligned} \tag{4.35}$$

□

Once again, the proof for Theorem 6 does not depend on the sample size used to develop the fiducial interval, and therefore the minimum desired coverage of $(1 - \alpha)100\%$ will be met for any sample size. Also, using the definition of the confidence coefficient,

$$Pr(BC_L < BC < BC_U) = Pr(Y < y \mid BC_L) - Pr(Y \leq y \mid BC_U) \tag{4.36}$$

the confidence coefficient for this CI for any $BC = \mathbf{p}^T\mathbf{1}$ can be calculated. However, the specific $\mathbf{p}$ must be known in order to determine the probability of observing each $\mathbf{X}$ in the $\mathcal{A}$ sample space. For this reason, the confidence coefficient can be calculated for a specific set of misclassification probabilities for each class, but not explicitly for a given BC, because there are infinite $\mathbf{p}$ that could result in each BC (except the trivial cases where $BC = 0$ or $BC = \sup\{\mathcal{BC}\}$).

4.2.3 Fiducial Interval around Bayes Cost Algorithm.

A general procedure is presented for finding the fiducial interval around BC in Section 4.2.3.1. A simplified procedure is presented in Section 4.2.3.2 for scenarios where the weights on misclassiifcation outcomes ($c_{i|j}p_j$) and all class sample sizes (n_j) are equal. If $\widehat{BC} = 0$ or $\widehat{BC} = \sup\{\mathcal{Y}\}$, the lower bound is 0 or the upper bound is $\sup\{\mathcal{BC}\}$, respectively. For such a case, the algorithm should be used to find the remaining upper or lower bound only.

4.2.3.1 General Case.

The following is an outline of steps to compute the proposed fiducial interval for k classes, an observed $y = \widehat{BC}$, and classification system with either equal or unequal weights (explained with options for equal [unequal] weights throughout).

1. Create the joint binomial [multinomial] sample space, $\mathcal{B}$ [$\mathcal{A}$] , for the k independent binomial [multinomial] distributions from each class for equal [unequal] weights (this sample space will have $\prod_{j=1}^{k}(n_j+1)\left[\prod_{j=1}^{k}\begin{pmatrix} n_j+k-1 \\ n_j \end{pmatrix}\right]$ elements).

2. Order the sample space, $\mathcal{B}$ [$\mathcal{A}$] , by each element's resulting $y = \sum_{i=1,i\neq j}^{k}\sum_{j=1}^{k} c_{i|j}p_j\frac{x_{i|j}}{n_j}$.

3. Create the joint binomial [multinomial] parameter space, $\mathbf{p} = (p_{1^c|1},\ldots,p_{k^c|k})$ [$\mathbf{p} = (\mathbf{p_1},\ldots,\mathbf{p_k})$] , to search for BC_L and BC_U. This parameter space is infinite, therefore the search for the upper and lower bounds on BC will only consider all $\mathbf{p}$ generated by a specified step or precision, δ . (It is recommended to start with a larger δ , such as $\delta = 0.2$ and consider smaller δ while narrowing in on the solution to conserve code run time.)

4. For each element of the parameter space created in Step 3, apply Equation 4.16 [4.8] and sum and store the resulting $f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{p})$ from all elements of the $\mathcal{B}$ [$\mathcal{A}$] sample space whose corresponding y is less than or equal to $\widehat{BC}$.

5. Calculate $BC = \sum_{j=1}^{k} p_{j\mathrm{c}|j}$ [$BC = \sum_{i=1,i\neq j}^{k}\sum_{j=1}^{k} c_{i|j}p_j p_{i|j}$] for each element in the joint parameter space created in Step 3.

6. For each fixed BC resulting from the parameter space found in Step 5, determine and store the maximum value of the sum in Step 4 (this gives $F_Y^1(y \mid BC)$).

6a. For unequal weights only, create the step function in Equation 4.29. This is done by determining the maximum value from Step 6 for all $BC \in \mathcal{BC}$ values greater than or equal to each specific BC value. For each BC value this gives $F_Y^3(y \mid BC)$.

7. The upper bound (BC_U) is determined as the smallest BC whose maximum value from Step 6 [6a] is $\leq \alpha/2$.

8. For the lower bound, repeat Steps 4-6, however, instead of summing the elements of the binomial [multinomial] sample space where $y \leq \widehat{BC}$, sum the elements for which $y > \widehat{BC}$. Then, for each BC from the binomial [multinomial] parameter space, determine the maximum value resulting from this sum which gives $1 - F_Y^2(y \mid BC)$.

8a. For unequal weights only, the maximum value of $1 - F_Y^2(y \mid BC)$ for all

$BC \in \mathcal{BC}$ less than or equal to each fixed BC value is found which gives $F_Y^4(y \mid BC)$ (Equation 4.30).

9. The lower bound (BC_L) is determined as the largest BC where $1 - F_Y^2(y \mid BC)$ $[F_Y^4(y \mid BC)]$ is $\leq \alpha/2$.

10. Improve the precision of the solution by repeating Steps 3-9 iteratively, using parameter spaces with smaller δ values. Before applying Steps 4-9 reduce the joint binomial [multinomial] parameter space to be searched by only considering elements resulting in BC values which are $\pm 2\delta$ from the previous BC_L or BC_U for finding the lower or upper bound, respectively.

4.2.3.2 Special Case: Equal Sample Sizes and Weights.

When sample sizes (n_j) and all weights ($c_{i|j}p_j$) on misclassification outcomes within the classes are equal, the previous steps may be used, or more efficiently, the following may be used. For this special case, the fiducial interval around BC reduces to a multiple of the Clopper-Pearson CI around a binomial probability of success (where a success is defined as an incorrect classification). This is demonstrated WLOG assuming an equal weight of one for all misclassification probabilities. First, it is possible to determine the total misclassification probability from the entire classification system as $p_{mc} = \frac{\sum_j^k p_j c_{|j}}{k} = \frac{BC}{k}$. Let the total number of misclassifications from the classification system be the binomial random variable $X = X_{1c|1} + \cdots + X_{kc|k}$. Then the binomial probability for the total misclassification of the system is estimated by $\widehat{p}_{mc} = \frac{X_{1c|1}+\cdots+X_{kc|k}}{k \times n} = \frac{\widehat{BC}}{k}$, which can be written in terms of $\widehat{BC}$ due to the equal sample size and weights in each class ($n = n_j$). Therefore, using the $(1-\alpha)100\%$ Clopper-Pearson fiducial interval constructed around p_{mc} such that $p_{mc} \in [p_{mc,L}, p_{mc,U}]$, the $(1-\alpha)100\%$ fiducial interval around BC is $BC \in [k \times p_{mc,L}, k \times p_{mc,U}] = [BC_L, BC_U]$. From this result, the fiducial interval for BC is easily computed as a multiple of the closed form solution to the Clopper-Pearson CI as

$$k \times \left[1 + \frac{N - x + 1}{xF_{2x,2(N-x+1),1-\alpha/2}}\right]^{-1} < BC < k \times \left[1 + \frac{N - x}{(x+1)F_{2(x+1),2(N-x),\alpha/2}}\right]^{-1} \tag{4.37}$$

where x is the total number of incorrect classifications observed for the entire sample from the classification system, F represents the F distribution, and $N = k \times n$ [12].

4.2.4 Equivalence for the Youden Index.

The fiducial interval around BC when all $c_{i|j}p_j$ are equal, for $i \neq j$, can be used equivalently for any k-class J. Because the outcomes from each class are modeled as binomial random variables in this framework, let the correct classifications from each class ($X_{j|j}$) be considered a success instead of the misclassifications ($X_{j^c|j}$). Then, the correct classification probability space ($\mathbf{p} = (p_{1|1}, \ldots, p_{k|k})$) will be searched for the upper and lower bounds. Now let $W = \widehat{J}$. Then

$$W = \sum_{\substack{i=1 \\ i=j}}^{k} \sum_{j=1}^{k} \frac{X_{i|j}}{n_j} \tag{4.38}$$

where the maximization is excluded because it is assumed the classifier is applied at its optimal settings. The $(1-\alpha)100\%$ upper and lower fiducial bounds for J from an observed statistic y are:

$$J_L = \sup \left\{ J \in \mathcal{J} \text{ such that } 1 - \min_{\mathbf{p}:\mathbf{p}^T\mathbf{1}=J} \left[\sum_{t \leq w^*} \sum_{\substack{\mathbf{x}\in\mathcal{B} \\ w=t}} f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{p}) \right] \leq \frac{\alpha}{2} \right\} \tag{4.39}$$

$$J_U = \inf \left\{ J \in \mathcal{J} \text{ such that } \max_{\mathbf{p}:\mathbf{p}^T\mathbf{1}=J} \left[\sum_{t \leq w} \sum_{\substack{\mathbf{x}\in\mathcal{B} \\ w=t}} f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{p}) \right] \leq \frac{\alpha}{2} \right\} \tag{4.40}$$

where $\mathcal{J}$ is the J sample space such that $\mathcal{J} = \{J : J = \mathbf{p}^T\mathbf{1}, \mathbf{p} = (p_{1|1}, \ldots, p_{|k}), p_{i|j} \in [0,1]\}$, $\mathcal{B}$ is the joint binomial sample space which is the set of $1 \times k$ sized vectors such that $\mathcal{B} = \{\mathbf{x} = (x_{1|1}, \ldots, x_{k|k}) : x_{j|j} \in \mathbb{Z}^+, x_{j|j} \leq n_j\}$, $\mathcal{W} = \{w : w = \sum_{i=1, i=j}^{k} \sum_{j=1}^{k} \frac{x_{i|j}}{n_j}, \mathbf{x} \in \mathcal{B}\}$ and w^* is the ordered value of $W \in \mathcal{W}$ directly less than w. When $w = 0$ or $w = k$, the lower bound is $J_L = 0$ and the upper bound is $J_U = k$, respectively.

4.3 Bootstrap Methods

Bootstrap methods presented in Section 3.4 may be similarly applied here. For comparison to the newly developed nonparametric CI around BC, the BCa bootstrap CI will be used. The BCa bootstrap CI is a practical choice because this CI method is appropriate when the distribution of the parameter is skewed [11]. Recall that since BC is constructed by the minimization of multinomial probabilities, it is expected that this distribution may be skewed. This was observed in the results of the simulation in Section 3.5. The BCa CI also allows the skewness of the distribution to change with the varying parameter, which also might be expected for BC based on the results of Section 3.5

(Figure 3.1) [11]. Finally, the BCa bootstrap CI is used for nonparametric CI around BC because this CI method was shown to perform best for CI around BC in Chapter 3 (with BC estimated parametrically) and for CI around J in [56] (with J estimated empirically).

4.4 Simulation Results

A simulation study was conducted to demonstrate the performance of the proposed fiducial interval around BC. This method is ideal for small sample sizes, and therefore the simulations are run with various equal and unequal small sample size scenarios for both the two- and three-class BC. For clarity in this section, the two-class BC is denoted BC_2 and the three-class BC denoted BC_3. These are defined as

$$BC_2 = c_{2|1}p_1p_{2|1} + c_{1|2}p_2p_{1|2} \tag{4.41}$$

and

$$BC_3 = \sum_{\substack{i=1 \\ i\neq j}}^{3}\sum_{j=1}^{3} c_{i|j}p_jp_{i|j} \tag{4.42}$$

Multiple values of BC_2 and BC_3 are considered in order to demonstrate performance of the fiducial interval around BC under differing classification system performance. In addition to varying classification performance scenarios, both equal and unequal weights are considered. The unequal weights scenarios utilize the two unequal cost structures from the simulation in Chapter 3. Recall that these cost structures are $Cost_1 = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}$ and $Cost_2 = \begin{bmatrix} 0 & 2 & 5 \\ 1 & 0 & 3 \\ 1 & 3 & 0 \end{bmatrix}$ where $Cost = \begin{bmatrix} c_{1|1} & c_{1|2} & c_{1|3} \\ c_{2|1} & c_{2|2} & c_{2|3} \\ c_{3|1} & c_{3|2} & c_{3|3} \end{bmatrix}$. All prevalences are assumed equal ($p_j = \frac{1}{3}$).

Two distributional scenarios are considered. First, no distributional assumptions about the classification system are made. Then, comparisons are made against other CI techniques when the single feature used for classification is independently and normally distributed for each class. Each distributional scenario is discussed separately. Absolute bias of $\widehat{BC}$ is also presented. All simulation scenarios used 3000 simulation runs in R and $\alpha = 0.05$ [52].

4.4.1 Equal Costs.

A cost structure is assumed where $c_{i|j}p_j = 1$, for $i \neq j$. Under this framework, $BC_2 \in [0, 2]$ and $BC_3 \in [0, 3]$, where $BC_2 = 1$ and $BC_3 = 1.5$ reflect chance classification. The values of BC chosen to reflect a range of classification accuracy are $BC_2 = (0.6, 0.4, 0.2, 0.1)$ and $BC_3 =$

$(0.9, 0.6, 0.3, 0.15)$, such that each BC_3 value has the same average misclassification probability as a corresponding BC_2 value.

4.4.1.1 No Distributional Assumptions on the System.

Making no assumptions about the classification system's structure, multinomial random variables are randomly generated representing outcomes from a classification system's resulting contingency table (recall Tables 2.3 and 2.4). The misclassification probabilities are assumed to be equally distributed between all classes for each BC_2 or BC_3 value. The fiducial interval is constructed around BC separately for all 3000 simulation runs and the coverage probability and average length of the intervals calculated. Absolute bias of the estimated BC is also calculated.

The results are presented in Table 4.4. For all sample size and BC scenarios, the intervals perform well with coverage probabilities of at least 95%. Also, the average length of the interval decreases as the total sample size increases and as the classification performance improves (smaller BC). The absolute bias in the empirically estimated BC is higher for larger BC values and decreases as n_j increases, mimicking the trend of interval length. Absolute bias is higher for BC_3 (absolute bias $\in [0.056, 0.292]$) than for BC_2 (absolute bias $\in [0.044, 0.225]$) for equivalent n_j .

4.4.1.2 Normally Distributed Feature.

To compare the performance of the proposed fiducial interval to other available CI methods for BC, a classification system with a single feature that is independently and normally distributed for each class and a single threshold between each class (two thresholds for BC_3) is assumed. For all scenarios, the variance for each class is assumed equal to one and the means are varied to achieve the desired BC_2 or BC_3 value. These normal distribution parameters are listed in Table 4.5. The sample sizes considered are held consistent with those in Section 4.4.1.1.

For both the two- and three-class scenarios, three methods in addition to the fiducial interval are compared. The first method is a nonparametric BCa bootstrap CI. In [56], the BCa bootstrap CI is shown to have good coverage around J_2 for $n_j \geq 50$ when J_2 is estimated empirically. However, in [36], the BCa bootstrap is shown to perform well for slightly smaller sample sizes when J_3 is estimated parametrically (defining J as a function of the normal distribution parameters from the features). Recall, when all $c_{i|j}p_j$ are equal, for $i \neq j$, J and BC may be used equivalently where

Table 4.4: Simulation coverage probability and length for 95% fiducial intervals around BC for two and three classes when all $c_{i|j}p_j$ are equal, for $i \neq j$, making no distributional assumptions on the classification system

			$BC_2 =$	0.6		0.4		0.2		0.1	
# of Classes	n_1	n_2		Cov	Len	Cov	Len	Cov	Len	Cov	Len
$k = 2$	5	5		99.03	1.12	99.57	1.01	99.00	0.85	98.70	0.74
	6	9		96.00	0.92	97.47	0.82	98.93	0.68	98.27	0.58
	10	10		97.53	0.83	98.10	0.73	99.00	0.58	98.50	0.48
	12	18		95.60	0.67	96.00	0.60	98.77	0.47	99.33	0.38
	20	20		96.37	0.59	97.27	0.52	97.03	0.41	98.70	0.31
	22	28		95.47	0.51	95.90	0.46	96.33	0.35	98.23	0.27
	30	30		96.70	0.48	96.50	0.43	97.27	0.33	99.10	0.25
			$BC_3 =$	0.9		0.6		0.3		0.15	
# of Classes	n_1	n_2	n_3	Cov	Len	Cov	Len	Cov	Len	Cov	Len
$k = 3$	5	5	5	97.80	1.41	98.00	1.26	98.73	1.02	99.40	0.85
	4	6	10	96.53	1.25	98.27	1.12	98.27	0.91	98.97	0.78
	10	10	10	96.97	1.02	97.93	0.91	99.17	0.71	98.53	0.56
	8	12	20	95.67	0.91	96.77	0.82	98.63	0.64	99.00	0.52
	20	20	20	96.60	0.72	96.27	0.64	97.03	0.49	98.97	0.37
	24	16	30	95.17	0.67	95.37	0.59	98.03	0.47	99.23	0.36
	30	30	30	96.17	0.59	96.47	0.52	96.13	0.40	97.53	0.30

Cov - coverage probability; Len - length

$BC_2 = 1 - J_2$ and $BC_3 = 3 - J_3$ (where $J_2 = p_{1|1} + p_{2|2} - 1$ and $J_3 = p_{1|1} + p_{2|2} + p_{3|3}$). Therefore, two BCa bootstrap CIs are constructed around both BC_2 and BC_3, one utilizing an empirical and the other a parametric estimation of BC as described in [56] and [36] (denoted BCa_E and BCa_P, respectively). For both BCa CIs, 999 nonparametric bootstrap samples are used. In addition, the delta method CI (see Section 3.2) and the GCI (see Section 3.3) are also used for comparison to the fiducial interval. For the implementation of these CIs, the classifier is applied to the random samples from the normal distributions to construct the resulting contingency table (in the spirit of Table 2.4), and then the appropriate CI method is applied.

One additional method is also considered for comparisons of CIs around BC_2. Because this method was developed for the two-class framework only, it is not used in the simulation for BC_3.

Table 4.5: Normal distribution parameters used in fiducial interval simulation with each $\sigma_j = 1$

# of Classes	BC	μ_1	μ_2	μ_3
$k = 2$	0.6	0	1.049	-
	0.4	0	1.683	-
	0.2	0	2.563	-
	0.1	0	3.290	-
$k = 3$	0.9	-1.0	0	2.148
	0.6	-1.5	0	2.902
	0.3	-2.5	0	3.405
	0.15	-3.6	0	3.523

This final method is a nonparametric method which assumes there is a single threshold between the two classes, but makes no assumptions about the distribution of the feature. It is based on the Agrestti Coull CI for a binomial proportion and utilizes a bootstrap to determine the CI bounds (denoted NP) [79]. This method uses an estimation of J (easily modified for BC) given in Equation 2.33. Once again, since all weights are fixed to be equal, this CI method may be used equivalently for BC_2. The coverage and length of all CIs around BC_2 and BC_3 is determined by the 3000 simulation runs for the normally distributed feature. All simulations are run in R and the *boot* package is used for all bootstrapped CIs [10, 15, 52].

The results are presented in Table 4.6 for two classes and Table 4.7 for three classes (due to the poor performance of the NP CI, these results are in the Appendix, Section B.2). The proposed fiducial method meets or exceeds the desired coverage probability of 95% for all sample size and BC values considered. Also, similar to the simulation scenario which made no assumptions about the underlying distributions, as the total sample size increases and BC value decreases, the length of the fiducial interval decreases. Since lower BC values indicate a more accurate classification system, the proposed CI will perform best (when also considering length) for accurate systems.

Table 4.6: Simulation coverage probability and length for multiple methods' 95% CI around BC_2 for two classes with a normally distributed feature when all $c_{i|j}p_j$ are equal, for $i \neq j$.

			Fiducial		Delta		BCa_P		BCa_E		GCI	
n_1	n_2	BC_2	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len
5	5	0.6	100.0	1.10	86.07	0.80	86.10	0.67	71.90	1.32	96.93	0.77
		0.4	99.87	0.98	85.17	0.72	83.67	0.62	80.97	1.42	97.53	0.75
		0.2	98.90	0.80	82.00	0.53	78.93	0.43	64.30	1.16	98.37	0.62
		0.1	98.80	0.70	78.37	0.37	75.77	0.28	21.80	0.73	98.60	0.50
6	9	0.6	96.57	0.91	90.03	0.71	91.07	0.63	88.30	0.91	96.53	0.67
		0.4	98.80	0.81	88.80	0.63	89.23	0.59	88.07	0.93	97.07	0.64
		0.2	99.27	0.65	85.70	0.46	85.47	0.42	77.40	0.73	97.80	0.51
		0.1	99.00	0.56	82.47	0.31	83.33	0.28	41.17	0.46	98.00	0.39
10	10	0.6	98.60	0.82	91.73	0.61	93.33	0.57	84.26	0.76	96.06	0.59
		0.4	97.03	0.72	90.33	0.54	92.23	0.53	90.20	0.76	96.10	0.55
		0.2	98.97	0.57	87.50	0.39	89.60	0.38	88.07	0.57	96.60	0.42
		0.1	98.80	0.46	85.97	0.26	87.97	0.26	37.20	0.34	96.63	0.31
12	18	0.6	96.83	0.67	92.63	0.52	93.63	0.50	92.33	0.67	95.40	0.51
		0.4	96.10	0.59	91.57	0.46	93.17	0.46	94.07	0.64	95.37	0.46
		0.2	99.10	0.46	89.40	0.33	91.93	0.33	91.87	0.52	96.00	0.35
		0.1	99.60	0.36	87.37	0.22	90.97	0.23	84.03	0.34	96.13	0.25
20	20	0.6	95.63	0.59	92.13	0.45	93.47	0.44	89.63	0.59	94.33	0.44
		0.4	97.47	0.52	91.67	0.39	93.27	0.39	94.33	0.55	94.57	0.39
		0.2	96.17	0.40	90.37	0.29	92.33	0.29	94.07	0.44	94.63	0.29
		0.1	99.00	0.30	88.97	0.19	91.73	0.20	78.10	0.31	94.76	0.21
22	28	0.6	95.87	0.51	92.90	0.40	93.80	0.40	93.27	0.53	94.80	0.40
		0.4	96.00	0.45	92.30	0.36	93.60	0.36	95.20	0.48	95.07	0.36
		0.2	95.60	0.35	91.20	0.26	93.23	0.26	94.43	0.39	95.43	0.26
		0.1	98.57	0.26	89.23	0.17	92.77	0.18	91.37	0.29	95.63	0.18
30	30	0.6	97.13	0.48	94.37	0.37	94.93	0.37	90.57	0.49	95.70	0.37
		0.4	97.00	0.43	93.93	0.32	94.50	0.33	94.83	0.44	95.80	0.32
		0.2	97.37	0.33	93.33	0.24	93.47	0.24	95.93	0.36	95.47	0.24
		0.1	99.27	0.24	91.70	0.16	93.12	0.16	84.87	0.27	95.63	0.17

Cov - coverage probability; Len - length; BCa_P - bias corrected and accelerated/parametric estimate BCa_E - bias corrected and accelerated/empirical estimate; GCI - generalized confidence interval

Table 4.7: Simulation coverage probability and length for multiple methods' 95% CI around BC_3 for three classes with a normally distributed feature when all $c_{i|j}p_j$ are equal, for $i \neq j$.

				Fiducial		Delta		BCa_P		BCa_E		GCI	
n_1	n_2	n_3	BC_3	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len
5	5	5	2.1	98.50	1.37	89.97	0.91	80.80	0.69	74.03	1.00	97.53	0.89
			2.4	99.43	1.21	88.33	0.85	82.23	0.66	84.53	0.93	96.80	0.87
			2.7	99.27	0.94	86.37	0.67	79.17	0.49	62.80	0.66	95.00	0.80
			2.85	99.63	0.78	84.63	0.49	77.03	0.34	29.7	0.35	93.77	0.69
4	6	10	2.1	96.80	1.21	87.33	0.87	78.73	0.66	82.47	1.07	97.40	0.85
			2.4	99.30	1.07	87.33	0.82	80.87	0.63	84.97	0.99	97.70	0.84
			2.7	98.77	0.86	84.70	0.64	78.67	0.47	74.50	0.70	97.30	0.75
			2.85	99.40	0.74	84.23	0.44	79.67	0.33	53.27	0.40	95.57	0.63
10	10	10	2.1	98.60	1.01	92.47	0.68	89.73	0.60	89.87	0.92	96.33	0.66
			2.4	98.23	0.89	92.33	0.63	91.00	0.57	92.60	0.86	96.30	0.63
			2.7	99.40	0.68	91.33	0.49	98.97	0.44	90.97	0.68	95.97	0.53
			2.85	98.70	0.52	89.77	0.35	89.47	0.32	74.53	0.43	95.33	0.42
8	12	20	2.1	96.63	0.90	91.27	0.64	87.97	0.57	89.20	0.88	96.13	0.62
			2.4	95.20	0.80	91.23	0.61	89.57	0.55	91.80	0.84	96.47	0.61
			2.7	98.83	0.63	90.00	0.47	89.03	0.42	90.73	0.67	96.60	0.51
			2.85	99.47	0.50	89.50	0.32	88.73	0.29	86.67	0.43	95.03	0.38
20	20	20	2.1	97.80	0.72	93.43	0.49	92.20	0.46	92.33	0.69	95.57	0.48
			2.4	97.13	0.63	92.80	0.45	92.93	0.43	93.93	0.64	95.57	0.45
			2.7	96.80	0.48	92.03	0.35	92.03	0.34	94.50	0.52	95.03	0.37
			2.85	99.20	0.36	91.10	0.24	91.50	0.24	92.73	0.38	94.03	0.27
24	16	30	2.1	95.53	0.66	92.83	0.47	91.17	0.44	92.77	0.68	95.93	0.45
			2.4	95.63	0.59	92.87	0.45	91.87	0.43	94.93	0.63	95.10	0.45
			2.7	97.27	0.46	92.33	0.35	91.10	0.33	94.43	0.51	94.60	0.37
			2.85	99.33	0.35	90.93	0.25	91.13	0.23	91.63	0.38	93.57	0.28
30	30	30	2.1	97.37	0.59	94.33	0.40	92.67	0.39	94.93	0.57	94.57	0.40
			2.4	97.43	0.52	93.87	0.37	93.47	0.36	94.7	0.52	94.67	0.37
			2.7	97.00	0.40	93.73	0.29	93.13	0.28	95.00	0.42	94.53	0.30
			2.85	97.77	0.29	93.33	0.20	93.27	0.20	90.97	0.32	94.33	0.22

Cov - coverage probability; Len - length; BCa_P - bias corrected and accelerated/parametric estimate
BCa_E - bias corrected and accelerated/empirical estimate; GCI - generalized confidence interval

The only other CI that approaches the desired coverage probability is the GCI. The GCI has lengths that are on average 25% shorter than the fiducial intervals. However, the GCI is only appropriate for a classification system with a single feature that is independently and normally distributed for each class. The GCI always outperforms the delta method CI in coverage, which is also constructed on the assumption of a normally distributed feature (this was already observed in Section 3.5 for the small sample size scenario).

The NP CI performs poorly with respect to coverage for highly accurate classifiers ($BC = 0.1$) and for all BC values for $n_j \leq 20$ (see Appendix B.2). Therefore, this CI is not appropriate for a nonparametric small sample CI around BC. Both bootstrap CIs (with BC estimated either parametrically or empirically) perform poorly for small sample size scenarios (with the BCa_P CI outperforming the BCa_E CI). In general, the bootstrap BCa CI with a parametric estimate of BC performs very similar to the delta method CI in both length and coverage, as is also seen in Section 3.5. The BCa_E CI performs fairly well in coverage for $n_j \geq 20$, although the coverage drops for $BC = 0.1$. Also, as the coverage of the BCa_E CI gets close to the desired level ($\approx 90 - 95\%$), this CI's length becomes very similar to, and usually slightly worse than, the length of the fiducial interval. This suggests that for a nonparametric method that meets the desired coverage, it may not be possible to achieve shorter lengths than that of the fiducial interval.

The parametric estimate of BC (used for the delta, generalized, and BCa_P CIs) has the lowest absolute bias (absolute bias $\in [0.001, 0.209]$), which is expected because this estimate is based on the assumptions used for the simulation. The empirically estimated BC (used for the BCa_E and fiducial intervals) has larger bias (absolute bias $\in [0.007, 0.278]$) than the parametric estimates but similar bias as seen in the simulation that used multinomial random variables. Finally, the bias for the estimate of BC used for the NP CI increases significantly as BC_2 decreases (absolute bias $\in [0.066, 0.356]$). This trend in bias was also noted in [79] for J. This increase in bias for decreased BC_2 may contribute to the decreased coverage for this method at lower BC values.

4.4.2 Unequal Costs.

To ensure performance of the fiducial CI is not degraded when the costs of misclassification are not equal, two additional cost scenarios (the same cost structures considered for the parametric CIs

in Section 3.5) are considered for the three-class BC. In Section 3.5, the different cost structures do not have an impact on the CI performance for a normally distributed feature. For this reason the cost structure performance is demonstrated with the multinomial distribution only, as the performance with a normally distributed feature is not expected to differ from what is presented in Tables 4.6 and 4.7.

Additionally, only sample sizes up to $n_j = 20$ are considered as a result of intensive computational time when using the multinomial distributions. The same average total misclassification probabilities considered for the equal cost scenarios are used for this simulation, with the error probabilities being evenly distributed throughout the classes. This results in different BC_3 values, where $BC_{3,Cost_1} = (0.4, 0.27, 0.13, 0.07)$ and $BC_{3,Cost_2} = (0.75, 0.5, 0.25, 0.125)$. The coverage probability and length of the CIs are presented in Table 4.8. As expected, the CI is achieving a coverage probability of at least $1 - \alpha$. Notably, the CI for the unequal cost scenarios are more conservative with respect to coverage than the equal cost scenarios due to the step function required for finding the bounds. Finally, bias of the estimated BC is similar to the previous sections (absolute bias $\in [0.026, 0.137]$ for $Cost_1$ and absolute bias $\in [0.053, 0.274]$ for $Cost_2$).

4.5 Comparisons to Multinomial Methods

One simple solution for a CI around BC is the construction of simultaneous CIs around the multinomial probabilities resulting from the classification system, and then summing these bounds to calculate upper and lower bounds around BC:

$$BC_L = \sum_{\substack{i=1 \\ i \neq j}}^{k} \sum_{j=1}^{k} p_{i|j,L} \tag{4.43}$$

$$BC_U = \sum_{\substack{i=1 \\ i \neq j}}^{k} \sum_{j=1}^{k} p_{i|j,U} \tag{4.44}$$

where $[p_{i|j,L}, p_{i|j,U}]$ is the $(1-\alpha)100\%$ CI around $p_{i|j}$ found using a simultaneous CI method for the j^{th} class' multinomial probabilities. With k classes, k sets of simultaneous CIs will be needed which may require an adjustment for multiple comparisons to construct the $(1-\alpha)100\%$ CI around BC.

Table 4.8: Simulation coverage probability for 95% fiducial intervals around BC for three classes and two different cost structures making no assumptions on the classification system.

			BC_3 =	0.4		0.27		0.13		0.07	
$Cost_1$	n_1	n_2	n_3	Cov	Len	Cov	Len	Cov	Len	Cov	Len
	5	5	5	99.27	0.76	99.63	0.67	98.93	0.55	98.43	0.48
	4	6	10	98.93	0.70	99.40	0.61	99.63	0.50	99.50	0.42
	10	10	10	99.13	0.55	98.60	0.48	99.17	0.38	98.67	0.30
	8	12	20	98.50	0.52	98.93	0.46	99.13	0.37	99.53	0.30
	20	20	20	98.90	0.40	98.63	0.35	99.03	0.27	99.13	0.20
			BC_3 =	0.75		0.5		0.25		0.125	
$Cost_2$	n_1	n_2	n_3	Cov	Len	Cov	Len	Cov	Len	Cov	Len
	5	5	5	99.13	1.42	99.13	1.26	98.30	1.06	99.47	0.96
	4	6	10	98.53	1.12	98.20	0.98	97.90	0.80	98.50	0.67
	10	10	10	98.30	1.06	98.70	0.92	98.83	0.72	98.67	0.60
	8	12	20	97.87	0.82	97.87	0.71	98.13	0.55	98.27	0.43
	20	20	20	97.87	0.77	98.07	0.67	98.87	0.51	99.20	0.39

Cov - Coverage probability; Len - Length

The simultaneous CI methods for multinomial probabilities listed in Section 2.7.2.1 that may be used for producing a CI around BC are considered in this section. In [71] the performance of these methods is evaluated with respect to coverage probability. The Gold (1963) and Goodman (1965) methods have a minimum possible coverage probability of zero, which is not desirable [71]. The Queensberry and Hurst, Fitzpatrick and Scott, and Sison and Glaz methods all have minimum coverage probabilities greater than zero, although notably not greater than $1 - \alpha$ [71]. The three methods whose minimum coverage probability is greater than zero are considered for constructing a CI around BC.

The final method considered (although it is not a simultaneous CI for multinomial proportions) is the Clopper-Pearson CI for a binomial proportion (presented in Section 2.7.1.1). Under the assumption that all $c_{i|j}p_j$ are equal, for $i \neq j$, the total misclassification probability from each class may be modeled as a binomial proportion and therefore the Clopper-Pearson CI can be utilized. The Clopper-Pearson CI for binomial proportions has a minimum coverage probability of at least $1 - \alpha$ [3].

4.5.1 Simulation Results.

A simulation was conducted to compare the performance of the Clopper-Pearson, Fitzpatrck and Scott, Queensberry and Hurst, Sison and Glaz, Wald, and Log Wald methods when used to construct CIs around BC (Wald and Log Wald intervals are developed for BC in the Appendix, Section A.4). A three-class scenario with equal misclassification weights is assumed ($c_{i|j}p_j = 1$, $i \neq j$), and $n_j = 5$, 10, and 30 is considered. The coverage probability and length of the intervals over all values of BC (in increments of 0.01, allowing misclassification probabilities to be randomly assigned within all classes for each sample and fixed BC value) are determined using 10,000 simulation runs. Although some of the methods considered in this section require the construction of k sets of simultaneous CIs, an adjustment for multiple comparisons (such as the Bonferroni adjustment to α) is not made since these methods' resulting CIs around BC without an adjustment all have coverage above $1 - \alpha$. A Bonferroni adjustment would only increase the coverage and length of the interval, which is not desired for comparison. The results are presented in Figure 4.3[6].

The simultaneous CI methods do not perform well with respect to CI length (although coverage is met) and generally are so wide that the CI would be useless. Also, as expected due to the poor performance of the Wald CI on binomial proportions, the Wald and Log Wald methods do not meet the desired coverage, although they have shorter lengths. The performance of the fiducial interval is presented in Figure 4.3 for $k = 3$ and $n_j = 5$,10, and 30 with the red line. Notably, the fiducial method outperforms the simultaneous CI methods as it exceeds the desired coverage with much shorter lengths.

[6]Discontinuities in the plots at $BC = 1.0$ and $BC = 2.0$ occur due to a change in how the probabilities were randomly assigned, which was necessary to ensure the BC values reached the desired levels.

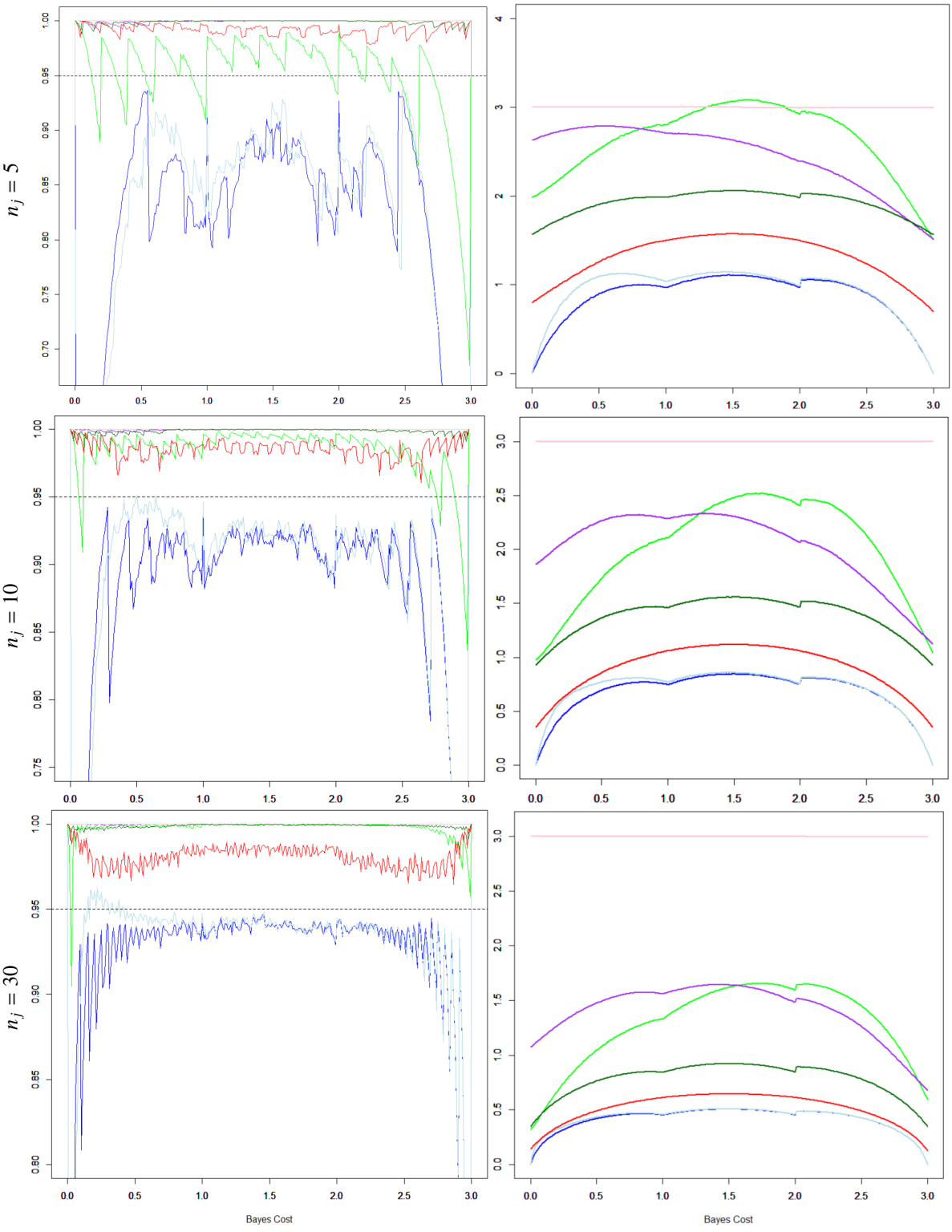

Figure 4.3: Coverage probability (left) and length (right) of CI methods for 95% CIs using Wald (blue), Log Wald (light blue), Sison and Glaz (green), Fitzpatrick and Scott (purple), Queensberry and Hurst (pink), Clopper and Pearson (dark green), and the fiducial interval developed in Section 4.2 (red).

4.6 Summary

Although Fisher's fiducial argument was lively debated, and deemed "Fisher's biggest blunder" by Efron, the objections to the theory were philosophical and not based on the method's feasibility [18, 31, 77]. In fact, in this chapter, the fiducial interval was shown to be a very useful and well performing tool for a CI around *BC*. The fiducial interval proposed in this chapter consistently meets the desired coverage probability for various classification scenarios. Although the CI has longer length than other intervals, when a CI under similar frameworks (empirically estimated BCa bootstrap CI) comes close to the desired coverage, the length of the other CI is similar and sometimes worse than that of the fiducial interval. The fiducial interval was shown to outperform the Wald, log Wald, and all simultaneous CI methods for multinomial probabilities considered with respect to coverage probability and length.

The fiducial interval performs well under any distributional scenario, as demonstrated in the simulation section using classification systems with either no underlying distributions or those with a single normally distributed feature. When the feature is normally distributed, the GCI presented in Section 3.3 outperforms all other methods considered in length, when the coverage was met. Coverage was met with both the GCI and fiducial methods, although the estimates of coverage were slightly lower with the GCI. The simulation suggests that the GCI performance may drop as class size and classification accuracy increases, in the three-class scenario. However, under the scenarios considered in the simulation in Section 4.4.1.2 for a normally distributed feature, the GCI is recommended. The utility of this CI is limited, however, as it is only appropriate for classification systems known to have thresholds between a feature's normal distributions for each class.

The fiducial interval has been developed to assure coverage is met. As such, the interval exceeds the coverage, resulting in interval lengths which may be seen as impractical. This is especially true for very small samples in the simulation. However, the fiducial interval is the only method which will guarantee coverage for any nonparametric scenario and sample size, and still may provide useful information. For instance, in the simulation where each class only has a sample of size five, coverage is about 99% for all *BC* values considered. For the high *BC* values, the lengths cover more than half of the possible range of *BC*. Yet when the *BC* value is low, the lengths of the

fiducial intervals are shorter to the extent that a classification system performing better than chance would still be determined. Therefore, even in small sample scenarios accurate systems may be detected, suggesting usefulness in this method for, say, pilot studies of potential classifiers.

The fiducial method requires searching the parameter space incremented by a predetermined tolerance. Given the step functions required for finding the bounds for BC when costs on misclassifications are unequal, this tolerance should be chosen carefully. If the space is searched too coarsely, the upper or lower bound may be found to be too small or large, respectively. This is demonstrated for a three-class example where the second cost structure ($Cost_2$) is used. The top plot in Figure 4.4 is the minimum coverage at all BC values when the solution to the bounds was found by searching the parameter space, incremented by 0.05. The bottom plot in Figure 4.4 is the minimum coverage at all BC values when the parameter space searched was incremented by 0.01. It is clear that for a specific scenario, the minimum coverage achieved was below the desired level of 95% when the space was searched too coarsely. This minimum coverage is improved however, for the more finely searched interval. Therefore, although the developed fiducial interval theoretically guarantees a coverage of $(1-\alpha)100\%$, the increment used for searching the parameter space must be chosen carefully for the practical implementation of the interval when costs are unequal.

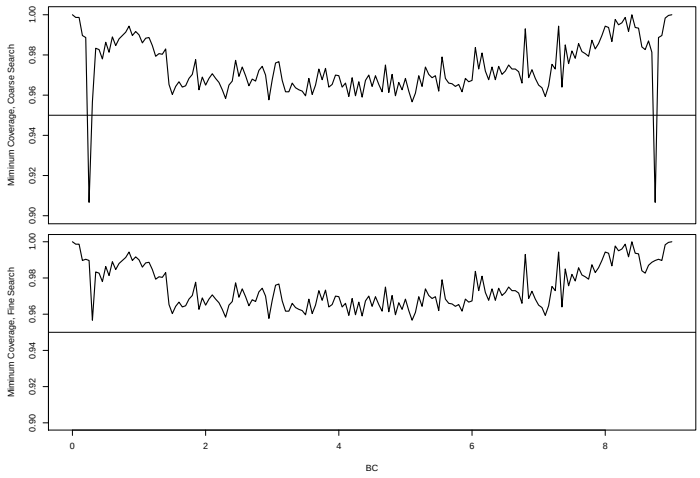

Figure 4.4: Minimum coverage of fiducial intervals when searching with coarser and finer increments in probability space, $\delta = 0.05$ (Top) and $\delta = 0.01$ (Bottom)

This chapter provides a nonparametric CI for any k-class BC which does not rely on information about the classification system used to construct the interval. This CI may be applied once the optimal thresholds have been selected, has the advantage of working for any classifier and regardless of scenario, and achieves the desired coverage probability. Therefore, in situations with small sample sizes or where the underlying distributions of the feature for each class are not normal or unknown, this fiducial interval provides a very useful and flexible tool for quantifying the uncertainty in $\widehat{BC}$. Finally, although this method is developed for applications with BC, it may be used for constructing a CI around any linear combination of multinomial or binomial probabilities.

V. Parametric Hypothesis Tests

5.1 Introduction

The methods proposed in this chapter assume a classification system with a single feature that is independently and normally distributed for each class and a threshold between ordered classes to distinguish any k number of classes. Under this framework, recall that BC can be written with the normal CDF, where the minimization is left off because this is achieved by using the $k-1$ optimal thresholds (θ_m^* , $m = 1, \ldots, k-1$).

$$BC = \sum_{j=2}^{k} c_{1|j} p_j \Phi\left(\frac{\theta_1^* - \mu_j}{\sigma_j}\right) + \sum_{\substack{i=2 \\ i \neq j}}^{k-1} \sum_{j=1}^{k} c_{i|j} p_j \left[\Phi\left(\frac{\theta_{m=i}^* - \mu_j}{\sigma_j}\right) - \Phi\left(\frac{\theta_{m=i-1}^* - \mu_j}{\sigma_j}\right)\right] + \sum_{j=1}^{k-1} c_{k|j} p_j \Phi\left(\frac{\mu_j - \theta_{k-1}^*}{\sigma_j}\right) \quad (3.9)$$

The development of two different types of hypothesis tests is considered. First, for a single classification system, it may be of interest to test a one sided hypothesis on BC in order to determine if the system performs at least as well as some pre-specified classification accuracy level (measured by BC). For instance, one may be interested in determining if a system performs better than chance. Lower values of BC correspond to better classification accuracy resulting in hypotheses of the form

$$H_0 : BC \geq BC_0 \text{ vs. } H_1 : BC < BC_0 \quad (5.1)$$

Secondly, it may also be of interest to compare the resulting BC values from two competing classification systems at their optimal point, in order to determine if one has superior classification performance. This hypothesis test may be of greater interest to decision makers because it provides information useful for choosing a classification system without having to specify a BC threshold (BC_0). It is assumed both classification systems are independent and have the same number of classes, and the feature used for each classification system is independently and normally distributed for each class. The two classification systems being compared will be denoted classification system A and classification system B. Define the difference between the two BC values from these systems as

$$\eta = BC_A - BC_B \quad (5.2)$$

The hypothesis to compare their performance would be of the form

$$H_0 : \eta \leq \eta_0 \text{ vs. } H_1 : \eta > \eta_0 \tag{5.3}$$

This hypothesis may be written for the specific case

$$H_0 : BC_A \leq BC_B \text{ vs. } H_1 : BC_A > BC_B \tag{5.4}$$

which is equivalent to testing at $\eta_0 = 0$ (no difference between performance). Higher values of BC indicate a classification system with poor performance and therefore the alternate hypothesis reflects the case when classification system B performs better than classification system A.

In Section 5.2, the delta method is used for developing both types of hypothesis tests assuming large sample sizes. In Section 5.3, both hypothesis tests are developed using a generalized hypothesis method for any sample size. A simulation is also considered to demonstrate the performance of the proposed hypothesis test methods (with size and power) and these results are presented in Section 5.4. Finally, a summary of the findings is presented in Section 5.5.

5.2 Delta Method Hypothesis Tests

5.2.1 One-sided Hypothesis Test on a Single Bayes Cost Value.

Recall from Section 3.2, for a normally distributed feature, as $n \to \infty$, $\widehat{BC} \sim N(BC, Var(\widehat{BC}))$ and the variance of $\widehat{BC}$ is estimated via the delta method with

$$Var\left(\widehat{BC}\right) \approx \sum_{j=1}^{k} \left[\left(\frac{\partial BC}{\partial \mu_j} \right)^2 Var(\widehat{\mu_j}) + \left(\frac{\partial BC}{\partial \sigma j} \right)^2 Var(\widehat{\sigma_j}) \right] \tag{3.10}$$

The partial derivatives for the three- and four-class BC are presented in Section 3.2.1 and Appendix A.3, respectively. However, for any number of classes, the partial derivatives are easily estimated numerically with the two point central difference method (Section 3.2.3). After estimating the partial derivatives and the variance of $\widehat{BC}$, the one sided hypothesis

$$H_0 : BC \geq BC_0 \text{ vs. } H_1 : BC < BC_0 \tag{5.1}$$

is tested by calculating a p-value from the observed sample. The p-value is developed using the following theorem.

Theorem 7 (Valid P-value)**.**

Let $W(\mathbf{X})$ be a test statistic such that large values of W give evidence that H_1 is true. For each sample point $\mathbf{x}$, define

$$p(\mathbf{x}) = \sup_{\theta \in \Theta_0} P_\theta(W(\mathbf{X}) \geq W(\mathbf{x}))$$

Then, $p(\mathbf{X})$ is a valid p-value. [12, p. 397]

Let $W(\mathbf{X}) = \frac{\widehat{BC} - BC_0}{\sqrt{Var(\widehat{BC})}}$. For $BC = BC_0$, $W(\mathbf{X})$ is distributed standard normal as $n \to \infty$. However, large values of $W(\mathbf{X})$ give evidence that H_1 is false. Therefore, to test the one-sided hypothesis in Equation 5.1 with this test statistic, the p-value is determined as

$$p(\mathbf{x}) = \sup_{\theta \in \Theta_0} P_\theta(W(\mathbf{X}) \leq W(\mathbf{x})) \tag{5.5}$$

For any arbitrary $BC' \geq BC_0$, $\widehat{BC} - BC_0 \geq \widehat{BC} - BC'$ and

$$P\left(Z \leq \frac{\widehat{BC} - BC_0}{\sqrt{Var(\widehat{BC})}}\right) \geq P\left(Z \leq \frac{\widehat{BC} - BC'}{\sqrt{Var(\widehat{BC})}}\right) \tag{5.6}$$

Therefore, the p-value for this hypothesis test is given by

$$\begin{aligned} p(\mathbf{x}) &= \sup_{\theta \in \Theta_0} P_\theta(W(\mathbf{X}) \leq W(\mathbf{x})) \\ &= P\left(Z \leq \frac{\widehat{BC} - BC_0}{\sqrt{Var(\widehat{BC})}}\right) \end{aligned} \tag{5.7}$$

At the α significance level, H_0 is rejected for $W(\mathbf{x}) \leq Z_\alpha$ or $p(\mathbf{x}) < \alpha$.

5.2.2 One-sided Hypothesis Test on the Difference of Two Bayes Cost Values.

For testing the hypothesis

$$H_0 : \eta \leq \eta_0 \text{ vs. } H_1 : \eta > \eta_0 \tag{5.3}$$

the parameter of interest, η , is a function of the normal distribution parameters, $\mu_{j,A}$, $\sigma_{j,A}$, $\mu_{j,B}$, and $\sigma_{j,B}$, $j = 1, \ldots, k$.

$$
\begin{aligned}
\eta =& BC_A - BC_B \\
=& \sum_{j=2}^{k} c_{1|j} p_j \Phi\left(\frac{\theta^*_{1,A} - \mu_{j,A}}{\sigma_{j,A}}\right) + \sum_{\substack{i=2 \\ i\neq j}}^{k-1} \sum_{j=1}^{k} c_{i|j} p_j \left[\Phi\left(\frac{\theta^*_{m=i,A} - \mu_{j,A}}{\sigma_{j,A}}\right) - \Phi\left(\frac{\theta^*_{m=i-1,A} - \mu_{j,A}}{\sigma_{j,A}}\right)\right] \\
&+ \sum_{j=1}^{k-1} c_{k|j} p_j \Phi\left(\frac{\mu_{j,A} - \theta^*_{k-1,A}}{\sigma_{j,A}}\right) - \sum_{j=2}^{k} c_{1|j} p_j \Phi\left(\frac{\theta^*_{1,B} - \mu_{j,B}}{\sigma_{j,B}}\right) \\
&- \sum_{\substack{i=2 \\ i\neq j}}^{k-1} \sum_{j=1}^{k} c_{i|j} p_j \left[\Phi\left(\frac{\theta^*_{m=i,B} - \mu_{j,B}}{\sigma_{j,B}}\right) - \Phi\left(\frac{\theta^*_{m=i-1,B} - \mu_{j,B}}{\sigma_{j,B}}\right)\right] \\
&- \sum_{j=1}^{k-1} c_{k|j} p_j \Phi\left(\frac{\mu_{j,B} - \theta^*_{k-1,B}}{\sigma_{j,B}}\right) \qquad (5.8)
\end{aligned}
$$

Therefore, from the multivariate delta method (Theorem 4), $\widehat{\eta} = g(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\sigma}})$ is Asymptotic-Normal$[\eta, Var(\widehat{\eta})]$ and $Var(\widehat{\eta})$ is estimated by

$$
\begin{aligned}
Var(\widehat{\eta}) \approx \sum_{j=1}^{k} \Bigg[& \left(\frac{\partial \eta}{\partial \mu_{j,A}}\right)^2 Var(\widehat{\mu_{j,A}}) + \left(\frac{\partial \eta}{\partial \sigma_{j,A}}\right)^2 Var(\widehat{\sigma_{j,A}}) \\
& + \left(\frac{\partial \eta}{\partial \mu_{j,B}}\right)^2 Var(\widehat{\mu_{j,B}}) + \left(\frac{\partial \eta}{\partial \sigma_{j,B}}\right)^2 Var(\widehat{\sigma_{j,B}}) \Bigg] \qquad (5.9)
\end{aligned}
$$

Covariance terms are excluded due to the assumption of independence between the normal distributions for each class and the classification systems being compared. Given that $\eta = BC_A - BC_B$ and each BC value only depends on the parameters associated with the classification system from which it were derived,

$$
\left(\frac{\partial \eta}{\partial \gamma_{j,A}}\right)^2 = \left(\frac{\partial BC_A}{\partial \gamma_{j,A}}\right)^2 \qquad (5.10)
$$

and

$$
\left(\frac{\partial \eta}{\partial \gamma_{j,B}}\right)^2 = \left(\frac{\partial BC_B}{\partial \gamma_{j,B}}\right)^2 \qquad (5.11)
$$

(where $\gamma_j = \mu_j$ or σ_j and $j = 1, \ldots, k$). Then Equation 5.9 may be rewritten:

$$Var(\widehat{\eta}) \approx \sum_{j=1}^{k} \left[\left(\frac{\partial BC_A}{\partial \mu_{j,A}} \right)^2 Var(\widehat{\mu_{j,A}}) + \left(\frac{\partial BC_A}{\partial \sigma_{j,A}} \right)^2 Var(\widehat{\sigma_{j,A}}) + \left(\frac{\partial BC_B}{\partial \mu_{j,B}} \right)^2 Var(\widehat{\mu_{j,B}}) + \left(\frac{\partial BC_B}{\partial \sigma_{j,B}} \right)^2 Var(\widehat{\sigma_{j,B}}) \right] \tag{5.12}$$

The partial derivatives required for estimating the variance of η in Equation 5.12 are found using the partial derivatives of BC. Recall these equations are presented in Section 3.2.1 for three classes, in Appendix A.3 for four classes, or generally for any k classes with the two-point central difference method presented in Section 3.2.3.

Similar to the previous section, the test statistic is $W(\mathbf{X}) = \frac{\widehat{\eta} - \eta_0}{\sqrt{Var(\widehat{\eta})}}$, where large values of $W(\mathbf{X})$ give evidence that H_1 is true. For $\eta = \eta_0$, $W(\mathbf{X})$ is distributed standard normal as $n \to \infty$, and the p-value for this hypothesis test is

$$p(\mathbf{x}) = P\left(Z \geq \frac{\widehat{\eta} - \eta_0}{\sqrt{Var(\widehat{\eta})}} \right) \tag{5.13}$$

At the α significance level, H_0 is rejected for $W(\mathbf{x}) \geq Z_{1-\alpha}$ or $p(\mathbf{x}) < \alpha$.

5.3 Generalized Hypothesis Tests

Let $\zeta = (\theta, \delta)$ where θ is the parameter of interest and δ is a vector of nuisance parameters.

Definition 4 (Generalized Test Variable)**.**

> *A random variable of the form $T = T(\mathbf{X}; \mathbf{x}, \zeta)$ is said to be a generalized test variable if it has the following three properties:*
> ***Property 1**: $t_{obs} = t(\mathbf{x}; \mathbf{x}, \zeta)$ does not depend on unknown parameters.*
> ***Property 2**: When θ is specified, T has a probability distribution that is free of nuisance parameters.*
> ***Property 3**: For fixed $\mathbf{x}$ and δ , $Pr(T \leq t; \theta)$ is a monotonic function of θ for any given t. [73, p. 115]*

If T is a generalized test variable which is stochastically decreasing in θ , the generalized p-value for testing $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$ can be found as [73, p. 119]

$$p(\mathbf{x}) = Pr(T \leq t_{obs} \mid \theta = \theta_0) \tag{5.14}$$

5.3.1 *One-sided Hypothesis Test on a Single Bayes Cost Value.*

For testing the hypotheses of the form

$$H_0 : BC \geq BC_0 \text{ vs. } H_1 : BC < BC_0 \tag{5.1}$$

the parameter of interest is the k-class BC defined in Equation 3.9, which is a function of the nuisance parameters, μ_j and σ_j , $j = 1, \ldots, k$. Define $T = T(\mathbf{X}; \mathbf{x}, \zeta)$ as

$$T = R_{BC} - BC \tag{5.15}$$

where R_{BC} was defined previously as

$$\begin{aligned} R_{BC} = &\sum_{j=2}^{k} c_{1|j} p_j \Phi\left(\frac{R_{\theta_1^*} - R_{\mu_j}}{R_{\sigma_j}}\right) \\ &+ \sum_{\substack{i=2 \\ i \neq j}}^{k-1} \sum_{j=1}^{k} c_{i|j} p_j \left[\Phi\left(\frac{R_{\theta_{m=i}^*} - R_{\mu_j}}{R_{\sigma_j}}\right) - \Phi\left(\frac{R_{\theta_{m=i-1}^*} - R_{\mu_j}}{R_{\sigma_j}}\right)\right] + \sum_{j=1}^{k-1} c_{k|j} p_j \Phi\left(\frac{R_{\mu_j} - R_{\theta_{k-1}^*}}{R_{\sigma_j}}\right) \end{aligned} \tag{3.25}$$

It was shown in Section 3.3.3, that R_{BC} is free of unknown parameters. Also recall that the GPQs for the optimal thresholds ($R_{\theta_m^*}$) are found numerically (when all $c_{i|j} p_j$ are not equal, for $i \neq j$). As seen in Section 3.3.1, for each class (indexed on $j = 1, \ldots, k$)

$$R_{\mu_j} = \bar{x}_j - t_j \frac{s_j}{\sqrt{n_j}} \tag{3.16}$$

and

$$R_{\sigma_j} = \sqrt{\frac{(n_j - 1)s_j^2}{V_j}} \tag{3.17}$$

where

$$t_j = \frac{\overline{X}_j - \mu_j}{S_j / \sqrt{n_j}} \tag{3.18}$$

and

$$V_j = \frac{(n_j - 1)S_j^2}{\sigma_j^2} \tag{3.19}$$

where $t_j \sim t_{(n_j - 1)}$, a t-distribution random variable with $n_j - 1$ degrees of freedom, and $V_j \sim \chi^2_{n_j - 1}$, a chi-square random variable with $n_j - 1$ degrees of freedom [12, p. 218, 223]. The observed value of T, where $t_{obs} = T(\bar{\mathbf{x}}, \mathbf{S})$, is evaluated by using $\bar{x}_j$ and S_j in Equations 3.18 and 3.19 and

then substituting Equations 3.18 and 3.19 into Equations 3.16 and 3.17, respectively. This results in $R_{\mu_j}(\bar{\mathbf{x}}, \mathbf{S}) = \mu_j$, $R_{\sigma_j}(\bar{\mathbf{x}}, \mathbf{S}) = \sigma_j$, and the numerically estimated $R_{\theta^*_m}(\bar{\mathbf{x}}, \mathbf{S}) = \theta^*_m$. Substituting $R_{\mu_j}(\bar{\mathbf{x}}, \mathbf{S}) = \mu_j$, $R_{\sigma_j}(\bar{\mathbf{x}}, \mathbf{S}) = \sigma_j$, and $R_{\theta^*_m}(\bar{\mathbf{x}}, \mathbf{S}) = \theta^*_m$ into Equation 3.25 and Equation 5.15 results in

$$\begin{aligned} t_{obs} =& \sum_{j=2}^{k} c_{1|j} p_j \Phi\left(\frac{\theta^*_1 - \mu_j}{\sigma_j}\right) \\ &+ \sum_{\substack{i=2 \\ i\neq j}}^{k-1} \sum_{j=1}^{k} c_{i|j} p_j \left[\Phi\left(\frac{\theta^*_{m=i} - \mu_j}{\sigma_j}\right) - \Phi\left(\frac{\theta^*_{m=i-1} - \mu_j}{\sigma_j}\right)\right] + \sum_{j=1}^{k-1} c_{k|j} p_j \Phi\left(\frac{\mu_j - \theta^*_{k-1}}{\sigma_j}\right) - BC \\ =& BC - BC \\ =& 0 \end{aligned} \tag{5.16}$$

Therefore, it is clear that Property 1 from Definition 4 is met since t_{obs} does not depend on unknown parameters. Property 2 of Definition 4 is met, because R_{BC} is free of unknown parameters which implies that when BC is specified, T does not depend on any nuisance parameters. Finally, for Property 3, let the distribution of R_{BC} be denoted $F_{R_{BC}}(r)$, which is free of unknown parameters. Since $T = R_{BC} - BC$, the CDF of T may be written as

$$\begin{aligned} Pr(T \leq t) =& Pr(R_{BC} \leq t + BC) \\ =& F_{R_{BC}}(t + BC) \end{aligned} \tag{5.17}$$

Therefore, BC is the location parameter for the distribution of T implying the CDF of T is a monotonic function of BC [12, pg. 116,134],[73, p. 117]. All three properties from Definition 4 are met for T defined in Equation 5.15 and therefore T is a generalized test variable which is stochastically decreasing in BC. From Equation 5.14 and 5.15, the generalized p-value for this test is given by

$$\begin{aligned} p(\mathbf{x}) =& Pr(T \geq t_{obs} \mid BC = BC_0) \\ =& Pr(R_{BC} - BC \geq t_{obs} \mid BC = BC_0) \\ =& Pr(R_{BC} - BC_0 \geq 0) \\ =& Pr(R_{BC} \geq BC_0) \end{aligned} \tag{5.18}$$

The probability in Equation 5.18 is evaluated via Monte Carlo methods by generating a large number of values for R_{BC} (in the same manner as was done in Section 3.3.3 for the GCI), and then determining the proportion of these values that satisfy the inequality in Equation 5.18 [73, p. 119].

5.3.2 One-sided Hypothesis Test on the Difference of Two Bayes Cost Values.

For testing the hypothesis

$$H_0 : \eta \leq \eta_0 \text{ vs. } H_1 : \eta > \eta_0 \tag{5.3}$$

recall the parameter of interest, η , is a function of the nuisance parameters, $\mu_{j,A}$, $\sigma_{j,A}$, $\mu_{j,B}$, and $\sigma_{j,B}$, $j = 1, \ldots, k$ (Equation 5.8). Now define $T = T(\mathbf{X}; \mathbf{x}, \zeta)$ as

$$T = R_\eta - \eta \tag{5.19}$$

where R_η is defined as

$$R_\eta = R_{BC_A} - R_{BC_B} \tag{5.20}$$

and R_{BC_A} and R_{BC_B} are defined as in Equation 3.25, by using Equations 3.16 through 3.19 with the appropriate sample mean, standard deviation, and sample size for each class within each system. It is clear following the same reasoning as was presented in Section 5.3.1, that $t_{obs} = 0$ and all three properties from Definition 4 are met for T in Equation 5.19. Thus, $T = R_\eta - \eta$ is a generalized test variable which is stochastically decreasing in η . From Equation 5.14 and 5.19, the generalized p-value for this test is

$$\begin{aligned} p(\mathbf{x}) =& Pr(T \leq t_{obs} \mid \eta = \eta_0) \\ =& Pr(R_\eta - \eta \leq t_{obs} \mid \eta = \eta_0) \\ =& Pr(R_\eta \leq \eta_0) \\ =& Pr(R_{BC_A} - R_{BC_B} \leq \eta_0) \end{aligned} \tag{5.21}$$

The probability in Equation 5.21 is evaluated via Monte Carlo methods by generating a large number of values for $R_{BC_A} - R_{BC_B}$ (in the same manner as was done in Section 3.3.3 for the GCI, however, now two *BC* GPQs are found, one for each classification system, and their difference stored). Then the proportion of these values that satisfy the inequality in Equation 5.21 is determined [73, p. 119].

5.4 Simulation Results

A simulation study was conducted to demonstrate the performance of the delta method and generalized hypothesis tests for BC and η . Various scenarios are considered including different sample sizes ($n_j = 10, 50, 100, 250$), underlying distributions of the feature used for classification (normal and gamma), differing costs associated with the misclassification outcomes, and classification accuracy (measured by BC/η value). All scenarios assume a classifier with three classes and two thresholds ($\theta_1^* < \theta_2^*$) to distinguish between adjacent classes.

All scenarios utilize 3000 simulation runs in R assuming a significance level of $\alpha = 0.05$. When required, numerical minimization is performed using the *optim* function in R ("L-BFGS-B" method) [52]. Performance of the hypothesis tests is measured with the simulation by estimating the size and power of each test.

Definition 5 (Power Function)**.**

> *The power function of a hypothesis test with rejection region R is the function of θ defined by $\beta(\theta) = P_\theta(\mathbf{X} \in R)$ [12, p. 383]*

Definition 6 (Size α Test)**.**

> *For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a size α test if $\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$.[12, p. 385]*

To evaluate the performance of the hypothesis test, the probability of rejecting the null hypothesis is determined for multiple BC (or η) values. The power function for a fixed sample size is monotone in θ (see for example, Figure 6.1). Therefore, $\beta(\theta)$ is first determined at the boundary of the null and alternate parameter space ($BC = BC_0$, $\eta = \eta_0$) to estimate the size of the test ($\sup_{\theta \in \Theta_0} \beta(\theta)$). Then values in the alternate hypothesis space ($BC < BC_0$, $\eta > \eta_0$) are evaluated to estimate the power at increasing increments within the alternate hypothesis. In Section 5.4.1, the performance of the one-sided hypothesis test on a single BC value is evaluated and in Section 5.4.2, the performance of the one-sided hypothesis test on the difference of two independent BC values is evaluated.

5.4.1 One-sided Hypothesis Test on a Single Bayes Cost Value.

Four BC_0 values are chosen to demonstrate a range of potential classification system performance thresholds. Under the assumption of all $c_{i|j}p_j = 1$, for $i \neq j$, $BC_0 = 0.3, 0.5, 1.0, 1.25$. For the two additional cost structures, chosen as the cost structures used in previous chapters,

BC_0 was chosen to reflect similar scenarios (ie. similar normal curves) with the appropriate cost structure applied (recall $Cost_1 = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}$, $Cost_2 = \begin{bmatrix} 0 & 2 & 5 \\ 1 & 0 & 3 \\ 1 & 3 & 0 \end{bmatrix}$, and all $p_j = \frac{1}{3}$). This results in $BC_{0,Cost1} = 0.1, 0.2, 0.35, 0.45$ and $BC_{0,Cost2} = 0.2, 0.4, 0.7, 0.9$. The normal distribution parameters to achieve these BC_0 values are presented in Table 5.1. To study the power at differing BC values in the alternate hypothesis, the means of the first and third classes are varied to achieve the required BC value.

Table 5.1: Distributional parameters for the parametric hypothesis test simulation.

Distribution	BC_0	Class 1		Class 2		Class 3	
Normal (Equal Costs)		μ	σ	μ	σ	μ	σ
	0.30	-2.879	1	0	1	2.879	1
	0.50	-2.301	1	0	1	2.301	1
	1.00	-1.349	1	0	1	1.349	1
	1.25	-0.978	1	0	1	0.978	1
Normal ($Cost_1$)		μ	σ	μ	σ	μ	σ
	0.10	-2.879	1	0	1	2.879	1
	0.20	-2.077	1	0	1	2.077	1
	0.35	-1.333	1	0	1	1.333	1
	0.45	-0.985	1	0	1	0.985	1
Normal ($Cost_2$)		μ	σ	μ	σ	μ	σ
	0.20	-2.976	1	0	1	2.976	1
	0.40	-2.187	1	0	1	2.187	1
	0.70	-1.408	1	0	1	1.408	1
	0.90	-0.989	1	0	1	0.989	1
Gamma (Equal Costs)		α	β	α	β	α	β
	0.30	1.3	1	2.3	3.7	5	10.743
	0.50	1.3	1	2.3	3.7	5	5.234
	1.00	1.3	1	2	1.5	4	1.889
	1.25	1.3	1	2	1.5	4	1.162

Gamma distributed features are also considered for the equal weights scenario in order to evaluate the performance of the hypothesis tests when the assumption of normality is not met. The gamma distributional parameters used are presented in Table 5.1. To vary the BC values for

evaluating the power of the test, the α and β parameters from the second and third classes are varied appropriately.

The size and power of the delta and generalized hypothesis tests for equal weights are presented in Table 5.2 for a normally distributed feature and in Table 5.5 for a gamma distributed feature[7]. Simulation results for a normally distributed feature with $Cost_1$ are presented in Table 5.3 and in Table 5.4 for $Cost_2$.

In general, the performance of the delta and generalized hypothesis tests are similar. Usually, the delta method hypothesis test is slightly more powerful than the generalized hypothesis test, however when this occurs the delta method test often has a size $> \alpha$, which is not desirable. Overall, the size of the generalized hypothesis tests is smaller than the size of the delta method hypothesis test, and is usually bounded $\leq \alpha$. For $n_j = 10$ and equal weights (Table 5.2), the delta method size far exceeds 0.05 ($\alpha \in [0.09, 0.118]$). Therefore, the generalized hypothesis test should be used over the delta method tests for small sample sizes to assure α is maintained. As BC_0 approaches the value of chance classification (BC=1.5 for a three-class scenario) the feature's distributions for each class become more overlapped, making determination of the optimal point and correct class ordering more difficult. Therefore, as BC_0 increases, the performance of both tests is degraded with respect to size (see Table 5.2, $BC_0 = 1.25$). This is more apparent when observing the generalized hypothesis test.

For the unequal cost scenarios with $n_j \geq 50$, the delta method performs better with respect to power if a size of $\approx \alpha$ is acceptable (Tables 5.3 and 5.4). However, the generalized hypothesis test has very similar power to the delta method test, and maintains size $\leq \alpha$ (except for the one case where $n_j = 10$ and $BC_0 = 0.9$ for $Cost_2$).

As expected, the performance of both methods is degraded when the feature is not normally distributed (see Table 5.5). Overall, the performance for the gamma distributed feature is fair for most scenarios and reflects the robustness in these methods for minor deviations from normality.

[7] A detectable difference equal to BC_0 would result in testing at $BC = 0$ which is not possible with a normal or gamma distributed feature. Instead, the power at this detectable difference is approximated by testing at $BC = 0.001$.

Table 5.2: Power for three classes with a normally distributed feature with equal weights. Detectable difference indicates the difference of the assumed true BC value and BC_0 ($BC < BC_0$). The power at a detectable difference of zero is the estimated size of the hypothesis test.

	Detectable	Delta Hypothesis Test				Generalized Hypothesis Test			
	Difference	n_j =10	50	100	250	10	50	100	250
$BC_0 = 0.30$	0 (α)	0.118	0.061	0.071	0.063	0.018	0.028	0.043	0.046
	0.01	0.131	0.086	0.106	0.122	0.023	0.038	0.069	0.092
	0.05	0.193	0.273	0.415	0.678	0.040	0.158	0.312	0.614
	0.10	0.332	0.660	0.882	0.999	0.087	0.508	0.823	0.997
	0.20	0.758	0.999	1.000	1.000	0.364	0.998	1.000	1.000
	0.30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$BC_0 = 0.50$	0 (α)	0.105	0.053	0.069	0.061	0.025	0.031	0.045	0.047
	0.01	0.114	0.073	0.092	0.103	0.030	0.040	0.067	0.082
	0.05	0.160	0.207	0.304	0.505	0.043	0.125	0.230	0.455
	0.10	0.239	0.465	0.695	0.957	0.075	0.340	0.617	0.944
	0.20	0.501	0.945	0.997	1.000	0.210	0.898	0.994	1.000
	0.30	0.810	1.000	1.000	1.000	0.503	1.000	1.000	1.000
	0.40	0.984	1.000	1.000	1.000	0.890	1.000	1.000	1.000
	0.50	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$BC_0 = 1.00$	0 (α)	0.090	0.055	0.064	0.056	0.044	0.043	0.054	0.051
	0.01	0.097	0.071	0.085	0.096	0.047	0.054	0.074	0.085
	0.05	0.134	0.164	0.246	0.399	0.069	0.128	0.213	0.383
	0.10	0.193	0.358	0.558	0.870	0.108	0.308	0.525	0.860
	0.20	0.380	0.818	0.975	1.000	0.229	0.788	0.968	1.000
	0.30	0.586	0.990	1.000	1.000	0.430	0.985	1.000	1.000
	0.40	0.798	1.000	1.000	1.000	0.649	1.000	1.000	1.000
	0.50	0.934	1.000	1.000	1.000	0.859	1.000	1.000	1.000
$BC_0 = 1.25$	0 (α)	0.095	0.062	0.061	0.057	0.068	0.054	0.055	0.054
	0.01	0.103	0.076	0.086	0.095	0.074	0.071	0.081	0.093
	0.05	0.144	0.165	0.240	0.401	0.106	0.153	0.231	0.394
	0.10	0.202	0.361	0.551	0.871	0.158	0.340	0.537	0.867
	0.20	0.384	0.813	0.971	1.000	0.315	0.798	0.970	1.000
	0.30	0.584	0.987	1.000	1.000	0.518	0.985	1.000	1.000
	0.40	0.789	1.000	1.000	1.000	0.725	1.000	1.000	1.000
	0.50	0.920	1.000	1.000	1.000	0.887	1.000	1.000	1.000

Table 5.3: Power for three classes with a normally distributed feature with the $Cost_1$ cost structure. Detectable difference indicates the difference of the assumed true BC value and BC_0 ($BC < BC_0$). The power at a detectable difference of zero is the estimated size of the hypothesis test.

	Detectable	Delta Hypothesis Test				Generalized Hypothesis Test			
	Difference	n_j =10	50	100	250	10	50	100	250
$BC_0 = 0.10$	0 (α)	0.118	0.061	0.071	0.063	0.017	0.027	0.043	0.046
	0.01	0.161	0.164	0.228	0.355	0.029	0.080	0.151	0.292
	0.05	0.526	0.951	0.997	1.000	0.175	0.891	0.994	1.000
	0.10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$BC_0 = 0.20$	0 (α)	0.097	0.052	0.068	0.061	0.022	0.030	0.045	0.046
	0.01	0.126	0.119	0.164	0.247	0.032	0.066	0.118	0.210
	0.05	0.318	0.698	0.917	0.999	0.116	0.590	0.880	0.999
	0.10	0.715	0.998	1.000	1.000	0.413	0.998	1.000	1.000
	0.15	0.980	1.000	1.000	1.000	0.892	1.000	1.000	1.000
	0.20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$BC_0 = 0.35$	0 (α)	0.089	0.062	0.064	0.060	0.032	0.038	0.046	0.049
	0.01	0.111	0.098	0.130	0.190	0.041	0.070	0.106	0.167
	0.05	0.252	0.519	0.764	0.980	0.109	0.436	0.712	0.974
	0.10	0.532	0.972	0.999	1.000	0.317	0.948	0.998	1.000
	0.15	0.840	1.000	1.000	1.000	0.636	1.000	1.000	1.000
	0.20	0.977	1.000	1.000	1.000	0.911	1.000	1.000	1.000
	0.25	1.000	1.000	1.000	1.000	0.996	1.000	1.000	1.000
	0.30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$BC_0 = 0.45$	0 (α)	0.106	0.055	0.071	0.062	0.038	0.044	0.048	0.051
	0.01	0.112	0.099	0.117	0.167	0.048	0.073	0.099	0.151
	0.05	0.236	0.437	0.660	0.939	0.117	0.375	0.616	0.931
	0.10	0.464	0.907	0.992	1.000	0.289	0.878	0.991	1.000
	0.15	0.742	0.998	1.000	1.000	0.556	0.997	1.000	1.000
	0.20	0.931	1.000	1.000	1.000	0.833	1.000	1.000	1.000
	0.25	0.993	1.000	1.000	1.000	0.965	1.000	1.000	1.000
	0.30	1.000	1.000	1.000	1.000	0.998	1.000	1.000	1.000

Table 5.4: Power for three classes with a normally distributed feature with the $Cost_2$ cost structure. Detectable difference indicates the difference of the assumed true BC value and BC_0 ($BC < BC_0$). The power at a detectable difference of zero is the estimated size of the hypothesis test.

	Detectable	Delta Hypothesis Test				Generalized Hypothesis Test			
	Difference	n_j =10	50	100	250	10	50	100	250
$BC_0 = 0.20$	0 (α)	0.136	0.072	0.077	0.065	0.022	0.034	0.047	0.046
	0.01	0.158	0.105	0.134	0.158	0.028	0.049	0.085	0.118
	0.05	0.274	0.415	0.604	0.890	0.066	0.269	0.497	0.846
	0.10	0.515	0.903	0.992	1.000	0.175	0.811	0.982	1.000
	0.15	0.828	1.000	1.000	1.000	0.454	1.000	1.000	1.000
	0.20	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000
$BC_0 = 0.40$	0 (α)	0.119	0.065	0.071	0.060	0.031	0.037	0.050	0.046
	0.01	0.132	0.084	0.113	0.119	0.036	0.054	0.077	0.099
	0.05	0.193	0.253	0.380	0.634	0.065	0.172	0.306	0.588
	0.10	0.310	0.600	0.829	0.993	0.115	0.487	0.773	0.991
	0.15	0.465	0.892	0.991	1.000	0.206	0.830	0.984	1.000
	0.20	0.656	0.992	1.000	1.000	0.339	0.983	1.000	1.000
	0.25	0.828	1.000	1.000	1.000	0.552	1.000	1.000	1.000
	0.30	0.954	1.000	1.000	1.000	0.784	1.000	1.000	1.000
$BC_0 = 0.70$	0 (α)	0.104	0.062	0.070	0.058	0.046	0.047	0.057	0.051
	0.01	0.114	0.080	0.095	0.102	0.049	0.060	0.077	0.091
	0.05	0.160	0.198	0.294	0.497	0.082	0.150	0.255	0.470
	0.10	0.243	0.460	0.669	0.951	0.128	0.392	0.632	0.946
	0.15	0.339	0.725	0.938	1.000	0.190	0.668	0.920	1.000
	0.20	0.468	0.921	0.997	1.000	0.285	0.885	0.994	1.000
	0.25	0.609	0.988	1.000	1.000	0.403	0.983	1.000	1.000
	0.30	0.731	0.999	1.000	1.000	0.549	0.999	1.000	1.000
$BC_0 = 0.90$	0 (α)	0.105	0.065	0.063	0.058	0.064	0.056	0.056	0.053
	0.01	0.113	0.082	0.094	0.100	0.072	0.073	0.086	0.095
	0.05	0.155	0.187	0.267	0.457	0.103	0.161	0.251	0.441
	0.10	0.226	0.417	0.622	0.915	0.151	0.386	0.601	0.910
	0.15	0.318	0.672	0.903	0.999	0.225	0.636	0.891	0.999
	0.20	0.427	0.870	0.990	1.000	0.314	0.845	0.987	1.000
	0.25	0.553	0.973	1.000	1.000	0.420	0.965	1.000	1.000
	0.30	0.670	0.997	1.000	1.000	0.545	0.996	1.000	1.000

Table 5.5: Power for three classes with a gamma distributed feature with equal weights. Detectable difference indicates the difference of the assumed true BC value and BC_0 ($BC < BC_0$). The power at a detectable difference of zero is the estimated size of the hypothesis test.

	Detectable	Delta Hypothesis Test				Generalized Hypothesis Test			
	Difference	n_j =10	50	100	250	10	50	100	250
$BC_0 = 0.30$	0 (α)	0.085	0.038	0.034	0.014	0.012	0.017	0.018	0.012
	0.01	0.092	0.051	0.050	0.031	0.015	0.023	0.032	0.022
	0.05	0.143	0.170	0.216	0.353	0.026	0.097	0.160	0.294
	0.10	0.273	0.573	0.817	0.990	0.087	0.457	0.756	0.986
	0.20	0.603	0.989	1.000	1.000	0.296	0.974	1.000	1.000
	0.30	0.832	0.989	0.999	1.000	0.975	1.000	1.000	1.000
$BC_0 = 0.50$	0 (α)	0.123	0.089	0.107	0.089	0.031	0.055	0.080	0.073
	0.01	0.131	0.111	0.139	0.153	0.036	0.070	0.111	0.130
	0.05	0.182	0.264	0.375	0.607	0.054	0.188	0.315	0.571
	0.10	0.270	0.547	0.768	0.975	0.093	0.432	0.711	0.968
	0.20	0.534	0.965	0.999	1.000	0.242	0.932	0.998	1.000
	0.30	0.747	1.000	1.000	1.000	0.599	1.000	1.000	1.000
	0.40	0.913	1.000	1.000	1.000	0.883	1.000	1.000	1.000
	0.50	0.833	0.989	0.999	1.000	1.000	1.000	1.000	1.000
$BC_0 = 1.00$	0 (α)	0.142	0.171	0.221	0.313	0.073	0.146	0.202	0.298
	0.01	0.154	0.201	0.281	0.418	0.083	0.174	0.256	0.406
	0.05	0.205	0.371	0.546	0.819	0.123	0.329	0.515	0.807
	0.10	0.287	0.609	0.825	0.989	0.182	0.574	0.804	0.987
	0.20	0.484	0.930	0.995	1.000	0.358	0.913	0.995	1.000
	0.30	0.660	0.997	1.000	1.000	0.562	0.996	1.000	1.000
	0.40	0.778	1.000	1.000	1.000	0.766	1.000	1.000	1.000
	0.50	0.962	1.000	1.000	1.000	0.919	1.000	1.000	1.000
$BC_0 = 1.25$	0 (α)	0.129	0.099	0.085	0.061	0.117	0.094	0.083	0.061
	0.01	0.143	0.121	0.125	0.112	0.127	0.117	0.117	0.110
	0.05	0.198	0.258	0.341	0.493	0.169	0.240	0.332	0.490
	0.10	0.286	0.511	0.701	0.931	0.249	0.490	0.692	0.930
	0.20	0.490	0.913	0.992	1.000	0.439	0.905	0.991	1.000
	0.30	0.690	0.999	1.000	1.000	0.661	0.998	1.000	1.000
	0.40	0.839	1.000	1.000	1.000	0.847	1.000	1.000	1.000
	0.50	0.921	1.000	1.000	1.000	0.953	1.000	1.000	1.000

5.4.2 *One-sided Hypothesis Test on the Difference of Two Bayes Cost Values.*

To evaluate the performance of the delta method and generalized hypothesis tests used for comparing the performance of two independent classification systems with respect to their BC value, η_0 is fixed at zero. All three cost structures considered previously are also used here: all $c_{i|j}p_j = 1$ (for $i \neq j$), $Cost_1 = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}$, and $Cost_2 = \begin{bmatrix} 0 & 2 & 5 \\ 1 & 0 & 3 \\ 1 & 3 & 0 \end{bmatrix}$ (all with $p_j = \frac{1}{3}$). The purpose of this hypothesis test is to compare two competing classification systems, and therefore, the cost structure placed on classification system A and classification system B are always the same. When the costs of misclassification are equal, normal and gamma distributed features are considered. For the unequal cost scenarios only normally distributed features are used. In order to evaluate the size and power of the test, the performance of classification system A is fixed ($BC_A = 0.80$ for equal costs, $BC_A = 0.50$ for $Cost_1$, and $BC_A = 0.70$ for $Cost_2$) and the performance of classification system B is varied to achieve the desired η values.

The power and size of each hypothesis test is estimated by simulation. The results for equal costs are presented in Table 5.6 for a normally distributed feature and in Table 5.7 for a gamma distributed feature. The results for $Cost_1$ and $Cost_2$ are presented in Tables 5.8 and 5.9, respectively. When all $c_{i|j}p_j = 1$, for $i \neq j$, the delta and generalized hypothesis tests perform similarly well. Again, for $n_j = 10$ the delta method hypothesis test has size greater than α ($\alpha \in [0.053, 0.061]$), however not by a large margin, and maintained equivalent or higher power than the generalized hypothesis test (Table 5.6). The gamma distributed feature does not degrade the performance of the hypothesis tests as much as when testing a single BC value (see Section 5.4.1). In fact, the performance with the gamma distributed feature is good, with size $\approx \alpha$ (Table 5.7). Since η is the difference of the BC values and therefore is a function of the difference of the distributions, this test statistic may be more similar to a normal distribution as compared to the one-sided test on a single BC value with a gamma distributed feature.

Similar to the one sided hypothesis tests on a single BC value, when costs are unequal, the delta method hypothesis test has slightly higher power than the generalized hypothesis test. However, again the delta method hypothesis also has slightly worse size than the generalized hypothesis test (see Tables 5.8 and 5.9).

Table 5.6: Power for three-class systems with normally distributed features with equal weights for testing $\eta \leq 0$. Detectable difference indicates the difference of the assumed true value of $BC_A - BC_B$ ($\eta \geq 0$). The power at a detectable difference of zero is the estimated size of the hypothesis test.

	Detectable	Delta Hypothesis Test				Generalized Hypothesis Test			
	Difference	n_j =10	50	100	250	10	50	100	250
$\eta_0 = 0$	0 (α)	0.061	0.052	0.043	0.052	0.047	0.048	0.041	0.051
	0.01	0.066	0.060	0.056	0.078	0.049	0.057	0.054	0.076
	0.05	0.086	0.112	0.147	0.254	0.066	0.109	0.142	0.253
	0.10	0.116	0.222	0.345	0.629	0.095	0.213	0.338	0.628
	0.15	0.155	0.374	0.593	0.908	0.125	0.368	0.588	0.906
	0.20	0.206	0.558	0.816	0.991	0.169	0.550	0.811	0.991
	0.25	0.268	0.733	0.939	1.000	0.222	0.724	0.939	1.000
	0.30	0.340	0.877	0.989	1.000	0.293	0.869	0.989	1.000
	0.35	0.424	0.951	0.998	1.000	0.364	0.948	0.998	1.000
	0.40	0.504	0.985	1.000	1.000	0.450	0.983	1.000	1.000
	0.45	0.594	0.995	1.000	1.000	0.541	0.995	1.000	1.000
	0.50	0.685	1.000	1.000	1.000	0.638	1.000	1.000	1.000

Table 5.7: Power for three-class systems with gamma distributed features with equal weights for testing $\eta \leq 0$. Detectable difference indicates the difference of the assumed true value of $BC_A - BC_B$ ($\eta \geq 0$). The power at a detectable difference of zero is the estimated size of the hypothesis test.

	Detectable	Delta Hypothesis Test				Generalized Hypothesis Test			
	Difference	n_j =10	50	100	250	10	50	100	250
$\eta_0 = 0$	0 (α)	0.070	0.059	0.066	0.059	0.035	0.055	0.062	0.059
	0.01	0.073	0.067	0.082	0.085	0.038	0.064	0.078	0.085
	0.05	0.094	0.124	0.166	0.242	0.054	0.117	0.165	0.240
	0.10	0.121	0.215	0.322	0.542	0.081	0.209	0.321	0.539
	0.15	0.152	0.334	0.502	0.808	0.112	0.328	0.499	0.812
	0.20	0.194	0.480	0.690	0.953	0.157	0.478	0.694	0.954
	0.25	0.266	0.684	0.901	0.999	0.202	0.670	0.896	0.999
	0.30	0.333	0.841	0.974	1.000	0.263	0.834	0.975	1.000
	0.35	0.409	0.934	0.996	1.000	0.337	0.929	0.996	1.000
	0.40	0.499	0.981	1.000	1.000	0.422	0.979	1.000	1.000
	0.45	0.586	0.996	1.000	1.000	0.519	0.996	1.000	1.000
	0.50	0.681	0.999	1.000	1.000	0.614	0.999	1.000	1.000

Table 5.8: Power for three-class systems with normally distributed features with the $Cost_1$ structure for testing $\eta \leq 0$. Detectable difference indicates the difference of the assumed true value of $BC_A - BC_B$ ($\eta \geq 0$). The power at a detectable difference of zero is the estimated size of the hypothesis test.

	Detectable	Delta Hypothesis Test				Generalized Hypothesis Test			
	Difference	n_j =10	50	100	250	10	50	100	250
$\eta_0 = 0$	0 (α)	0.060	0.050	0.044	0.051	0.037	0.042	0.052	0.053
	0.01	0.071	0.078	0.073	0.109	0.043	0.065	0.088	0.109
	0.05	0.126	0.241	0.366	0.667	0.081	0.241	0.366	0.675
	0.10	0.233	0.605	0.857	0.997	0.168	0.601	0.851	0.995
	0.15	0.378	0.909	0.994	1.000	0.313	0.895	0.994	1.000
	0.20	0.570	0.992	1.000	1.000	0.512	0.991	1.000	1.000
	0.25	0.752	1.000	1.000	1.000	0.704	1.000	1.000	1.000
	0.30	0.893	1.000	1.000	1.000	0.869	1.000	1.000	1.000
	0.35	0.971	1.000	1.000	1.000	0.962	1.000	1.000	1.000
	0.40	0.995	1.000	1.000	1.000	0.995	1.000	1.000	1.000
	0.45	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0.50	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 5.9: Power for three-class systems with normally distributed features with the $Cost_2$ structure for testing $\eta \leq 0$. Detectable difference indicates the difference of the assumed true value of $BC_A - BC_B$ ($\eta \geq 0$). The power at a detectable difference of zero is the estimated size of the hypothesis test.

	Detectable	Delta Hypothesis Test				Generalized Hypothesis Test			
	Difference	n_j =10	50	100	250	10	50	100	250
$\eta_0 = 0$	0 (α)	0.053	0.053	0.048	0.051	0.037	0.045	0.048	0.049
	0.01	0.059	0.061	0.064	0.079	0.040	0.055	0.065	0.075
	0.05	0.082	0.129	0.168	0.300	0.057	0.120	0.171	0.286
	0.10	0.124	0.266	0.412	0.728	0.087	0.263	0.403	0.731
	0.15	0.179	0.453	0.706	0.964	0.131	0.462	0.699	0.966
	0.20	0.242	0.669	0.903	0.998	0.186	0.664	0.896	0.998
	0.25	0.319	0.850	0.983	1.000	0.255	0.837	0.982	1.000
	0.30	0.424	0.944	0.998	1.000	0.343	0.937	0.999	1.000
	0.35	0.528	0.988	1.000	1.000	0.443	0.984	1.000	1.000
	0.40	0.629	0.999	1.000	1.000	0.548	0.997	1.000	1.000
	0.45	0.727	1.000	1.000	1.000	0.672	1.000	1.000	1.000
	0.50	0.822	1.000	1.000	1.000	0.784	1.000	1.000	1.000

5.5 Summary

Generalized and delta method hypothesis tests were developed for testing one sided hypotheses on a single *BC* value as well as the difference between two *BC* values for comparing independent competing classification systems. Both methods are developed assuming classification systems that use a single feature that is independently and normally distributed for each class. The performance of the proposed methods was demonstrated with simulations that evaluated the power and size of the tests. Varying scenarios as well as null hypothesis values were considered with the simulation.

In general, the generalized hypothesis test performed better and could be recommended for both forms of hypotheses (tests on BC_0 and η) and the various cost scenarios. Although, the delta method tests performed similar to the generalized tests and often had greater power, their size was sometimes greater than α which is not desirable. However, the delta method performance was improved for tests on η , which might be due to the increase in total sample size (when considering two classification systems instead of one) or the structure of the test statistic itself. For both methods,

the performance with respect to size was degraded for testing against the BC_0 value which was close to chance classification ($BC_0 = 1.5$). When the assumption of normality was not met, the performance of the hypothesis tests on a single BC_0 value was degraded. However, for testing the difference of two BC values, the performance of the tests remained fairly consistent. Therefore, it seems that when testing a hypothesis on η , the methods are more robust to departures from normality.

VI. Nonparametric Hypothesis Tests

6.1 Introduction

In this chapter, hypothesis tests for testing the performance of a classification system with BC are developed, making no assumptions about the classification system's underlying feature distributions or structure. Instead, inference methods are derived from the resulting classification outcomes from a classification system at a fixed $\boldsymbol{\theta} \in \Theta$, as was done for the nonparametric CIs derived in Chapter 4. Under this nonparametric framework, it is assumed that the classification system outcomes from each class may be modeled with independent multinomial distributions.

This chapter will consider the same two hypotheses that were developed in Chapter 5 under the parametric framework. The first hypothesis tests whether or not a classification system performs at least as well as a specified threshold value, BC_0 , where

$$H_0 : BC \geq BC_0 \text{ vs. } H_1 : BC < BC_0 \tag{5.1}$$

The second hypothesis considered compares two independent competing classification systems' performance. This is done by testing η , the difference in BC values from the two systems where

$$\eta = BC_A - BC_B \tag{5.2}$$

and

$$H_0 : \eta \leq \eta_0 \text{ vs. } H_1 : \eta > \eta_0 \tag{5.3}$$

For the specific case of testing if classification system B is performing better than classification system A, this hypothesis is tested at $\eta_0 = 0$.

In Section 6.2, exact small sample methods are developed for testing both hypothesis tests, using the fiducial theory developed in Section 4.2. In Section 6.3, nonparametric hypothesis tests for both hypotheses are developed for large sample sizes using likelihood ratio tests (LRTs). A simulation is conducted to demonstrate the performance of the tests with respect to power and size, and the results are presented in Section 6.4. The overall findings and conclusions are presented in Section 6.5.

6.2 Exact Hypothesis Tests

Under the nonparametric framework and when sample sizes are small, hypothesis tests may be conducted using similar exact methods as those used for developing fiducial intervals around BC in Section 4.2. In fact, the fiducial intervals presented in Section 4.2 are simply the inversion of the acceptance region of a two sided hypothesis test ($BC = BC_0$) on BC [7].

6.2.1 One-sided Hypothesis Test on a Single Bayes Cost Value.

The hypothesis of the form

$$H_0 : BC \geq BC_0 \text{ vs. } H_1 : BC < BC_0 \tag{5.1}$$

may be tested by calculating an associated p-value for the test. Recall from Theorem 7 (Section 5.2.1) that a valid p-value is given by

$$p(\mathbf{x}) = \sup_{\theta \in \Theta_0} P_\theta(W(\mathbf{X}) \geq W(\mathbf{x})) \tag{6.1}$$

when large values of $W(\mathbf{X})$ give evidence that H_1 is true. For the nonparametric framework, $W(\mathbf{X})$ is $\widehat{BC}$ defined empirically as

$$Y = \widehat{BC} = \sum_{\substack{i=1 \\ i \neq j}}^{k} \sum_{j=1}^{k} c_{i|j} p_j \frac{X_{i|j}}{n_j} \tag{4.27}$$

where each $X_{i|j}$ represents the number of observations classified as the i^{th} class when their true class is j , n_j is the total number of observations for the j^{th} class, and each $X_{i|j}$ is distributed multinomial. Once again for the hypothesis in Equation 5.1, large values of $W(\mathbf{X})$ give evidence that H_1 is false, and therefore the p-value for this test is

$$p(\mathbf{x}) = \sup_{\theta \in \Theta_0} P_\theta(W(\mathbf{X}) \leq W(\mathbf{x})) \tag{5.5}$$

Under the multinomial framework, a restriction on the BC parameter space is also a restriction on the joint multinomial parameter space, $\mathcal{S} = \{\mathbf{p} = (\mathbf{p_1}, \ldots, \mathbf{p_k}) : \mathbf{p_j} = (p_{1|j}, \ldots, p_{k|j}), p_{i|j} \geq 0$, and $\sum_{i=1}^{k} p_{i|j} = 1\}$. Thus, the hypotheses may be rewritten as

$$H_0 : \mathbf{p} \in \mathcal{S}_0 \text{ vs. } H_1 : \mathbf{p} \in \mathcal{S}_0^C \tag{6.2}$$

where $\mathcal{S}_0$ is the set of multinomial probabilities which result in $BC \geq BC_0$, and is defined as $\mathcal{S}_0 = \{\mathbf{p} = (\mathbf{p_1}, \ldots, \mathbf{p_k}) : \mathbf{p_j} = (p_{1|j}, \ldots, p_{k|j}), p_{i|j} \geq 0, \sum_{i=1}^{k} p_{i|j} = 1,$ and $\sum_{\substack{i=1 \\ i \neq j}}^{k} \sum_{j=1}^{k} c_{i|j} p_j p_{i|j} \geq BC_0\}$.

From an observed $\widehat{BC}$, the exact p-value for testing the hypothesis $BC \geq BC_0$ is given by

$$\begin{aligned} p(\mathbf{x}) &= \sup_{BC \geq BC_0} P_{BC}(Y \leq y) \\ &= \sup_{\boldsymbol{p} \in \mathcal{S}_0} P_{\boldsymbol{p}}(Y \leq y) \\ &= \sup_{\boldsymbol{p} \in \mathcal{S}_0} \sum_{t=0}^{y} \sum_{\substack{\mathbf{x} \in \mathcal{A} \\ y=t}} f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{p}) \end{aligned} \tag{6.3}$$

where $\mathcal{A}$ is the joint multinomial sample space which is the set of $1 \times k^2$ sized vectors $\mathbf{x} = (x_{1|1}, x_{2|1}, \ldots, x_{k-1|k}, x_{k|k})$ where each $x_{i|j}$ is a nonnegative integer and $\sum_{i=1}^{k} x_{i|j} = n_j$, $\mathbf{p} \in \mathcal{S}$, and

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{p}) &= \prod_{j=1}^{k} f_{\mathbf{X_j}}(\mathbf{x_j}) \\ &= \prod_{i=1}^{k} \prod_{j=1}^{k} n_j! \frac{p_{i|j}^{x_{i|j}}}{x_{i|j}!} \end{aligned} \tag{4.8}$$

The hypothesis is tested by calculating the p-value in Equation 6.3 and comparing this value to the chosen significance level, α . For $p(\mathbf{x})$ less than α , the null hypothesis is rejected.

6.2.2 One-sided Hypothesis Test on the Difference of Two Bayes Cost Values.

To test the hypothesis of the form

$$H_0 : \eta \leq \eta_0 \text{ vs. } H_1 : \eta > \eta_0 \tag{5.3}$$

under the framework of an exact hypothesis test, modeling the outcomes from the two classification systems with independent multinomial distributions, the parameter of interest, η , is a function of multinomial probabilities such that

$$\eta = \sum_{i=1, i \neq j}^{k} \sum_{j=1}^{k} c_{i|j,A} p_j p_{i|j,A} - \sum_{i=1, i \neq j}^{k} \sum_{j=1}^{k} c_{i|j,B} p_j p_{i|j,B} \tag{6.4}$$

and

$$Y = \widehat{\eta} = \sum_{i=1, i \neq j}^{k} \sum_{j=1}^{k} c_{i|j,A} p_j \frac{X_{i|j,A}}{n_{j,A}} - \sum_{i=1, i \neq j}^{k} \sum_{j=1}^{k} c_{i|j,B} p_j \frac{X_{i|j,B}}{n_{j,B}} \tag{6.5}$$

Define $\mathcal{A}_A$ as the joint multinomial sample space for classification system A, which is the set of $1 \times k^2$ sized vectors where $\mathcal{A}_A = \{\mathbf{x_A} = (\mathbf{x_{1,A}}, \ldots, \mathbf{x_{k,A}}) : \mathbf{x_{k,A}} = (x_{1|j,A}, \ldots, x_{k|j,A}), x_{i|j,A} \in \mathbb{Z}^+, \sum_{i=1}^{k} x_{i|j,A} = n_{j,A}\}$. Similarly, define $\mathcal{A}_B$ as the analogous joint multinomial sample space for classification system

B. Then the sample space for the entire experiment (for both classification systems) may be defined as $\mathcal{A}_{A,B}$ which is the set of $1 \times 2k^2$ sized vectors where $\mathcal{A}_{A,B} = \{(\mathbf{x_A}, \mathbf{x_B}) : \mathbf{x_A} \in \mathcal{A}_A, \mathbf{x_B} \in \mathcal{A}_B\}$.

Also, define the joint multinomial probability space for classification system A where $\mathbf{p_A} \in \mathcal{S} = \{\mathbf{p_A} = (\mathbf{p_{1,A}}, \ldots, \mathbf{p_{k,A}}) : \mathbf{p_{j,A}} = (p_{1|j,A}, \ldots, p_{k|j,A}), p_{i|j,A} \geq 0$, and $\sum_{i=1}^{k} p_{i|j,A} = 1\}$ and similarly define $\mathbf{p_B}$ for classification system B. The pmf for this experiment is the joint multinomial distribution from both classification systems such that

$$\begin{aligned} f_{\mathbf{X_A},\mathbf{X_B}}(\mathbf{x_A}, \mathbf{x_B} \mid \mathbf{p_A}, \mathbf{p_B}) &= \prod_{j=1}^{k} f_{\mathbf{X_{j,A}}}(\mathbf{x_{j,A}}) \times f_{\mathbf{X_{j,B}}}(\mathbf{x_{j,B}}) \\ &= \prod_{i=1}^{k} \prod_{j=1}^{k} n_{j,A}! \frac{p_{i|j,A}^{x_{i|j,A}}}{x_{i|j,A}!} n_{j,B}! \frac{p_{i|j,B}^{x_{i|j,B}}}{x_{i|j,B}!} \end{aligned} \tag{6.6}$$

Once again, the hypotheses may be rewritten as a restriction on the joint multinomial parameter space.

$$H_0 : (\mathbf{p_A}, \mathbf{p_B}) \in \mathcal{S}_0^2 \text{ vs. } H_1 : (\mathbf{p_A}, \mathbf{p_B}) \in \mathcal{S}_0^{2^C} \tag{6.7}$$

where $\mathcal{S}_0^2 = \{(\mathbf{p_A}, \mathbf{p_B}) : \mathbf{p_A} \in \mathcal{S}, \mathbf{p_B} \in \mathcal{S}$ and $\eta \leq \eta_0\}$. Then, for an observed $\widehat{\eta}$ from a classification system, the exact p-value for testing the hypothesis in Equation 5.3 is

$$\begin{aligned} p(\mathbf{x}) &= \sup_{\eta \leq \eta_0} P_\eta(Y \geq y) \\ &= \sup_{(\mathbf{p_A},\mathbf{p_B}) \in \mathcal{S}_0^2} P_{(\mathbf{p_A},\mathbf{p_B})}(Y \geq y) \\ &= \sup_{(\mathbf{p_A},\mathbf{p_B}) \in \mathcal{S}_0^2} \sum_{t=y}^{\sup\{\mathcal{Y}\}} \sum_{\substack{(\mathbf{x_A},\mathbf{x_B}) \in \mathcal{A}_{\mathcal{A},\mathcal{B}} \\ Y=t}} f_{\mathbf{X_A},\mathbf{X_B}}(\mathbf{x_A}, \mathbf{x_B} \mid \mathbf{p_A}, \mathbf{p_B}) \end{aligned} \tag{6.8}$$

where $\mathcal{Y} = \{y : y = \sum_{i=1, i\neq j}^{k} \sum_{j=1}^{k} c_{i|j,A} p_j \frac{X_{i|j,A}}{n_{j,A}} - \sum_{i=1, i\neq j}^{k} \sum_{j=1}^{k} c_{i|j,B} p_j \frac{X_{i|j,B}}{n_{j,B}}, (\mathbf{x_A}, \mathbf{x_B}) \in \mathcal{A}_{A,B}\}$. For an observed value of η , Y , and a fixed η_0 , the hypothesis is tested by calculating the p-value in Equation 6.8 and comparing this value to the chosen significance level, α . If $p(\mathbf{x})$ is less than α , reject the null hypothesis.

6.3 Likelihood Ratio Tests

LRTs are a general and common method that may be applied for hypothesis testing. Asymptotic properties of the likelihood ratio also make these tests easy to implement under large

sample assumptions. For the nonparametric methods developed in this section, it is assumed that each class has a large sample size ($n_j \gtrapprox 50$).

Definition 7 (Likelihood Ratio Test Statistic)**.**

The likelihood ratio test statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^C$ is

$$\lambda(x) = \frac{sup_{\Theta_0} L(\theta \mid x)}{sup_{\Theta} L(\theta \mid x)}$$

[12, p. 375]

To conduct a hypothesis test using the likelihood test statistic for large samples sizes, the following theorem may be used:

Theorem 8. *Let $X_1, \ldots, X_n$ be a random sample from a pdf or pmf $f(x \mid \theta)$. Under the regularity conditions . . . , if $\theta \in \Theta_0$, then the distribution of the statistic $-2 \log \lambda(\mathbf{X})$ converges to a chi squared distribution as the sample size $n \to \infty$. The degrees of freedom of the limiting distribution is the difference between the number of free parameters specified by $\theta \in \Theta_0$ and the number of free parameters specified by $\theta \in \Theta$ [12, p. 490].*

Regularity conditions are addressed in the Appendix, Section A.5.

6.3.1 One-sided Hypothesis Test on a Single Bayes Cost Value.

For this nonparametric large sample framework, it is again assumed that the outcomes from the classification system are distributed multinomial. Recall from Section 6.2, that under this framework, the one sided hypothesis on a single BC value may be written as a restriction on the joint multinomial parameter space:

$$H_0 : \mathbf{p} \in \mathcal{S}_0 \text{ vs. } H_1 : \mathbf{p} \in \mathcal{S}_0^C \tag{6.2}$$

The likelihood function is a function of the parameters, $\mathbf{p}$, with the data assumed given. Thus, the likelihood is comprised of the multinomial pmf, however it may be simplified by removing the constant multipliers which do not depend on the parameters. Therefore,

$$L(\mathbf{p} \mid \mathbf{x}) \propto \prod_{i=1}^{k} \prod_{j=1}^{k} p_{i|j}^{x_{i|j}} \tag{6.9}$$

An unrestricted maximization ($\sup_{\Theta} L(\mathbf{p} \mid \mathbf{x})$) of this likelihood results in the multinomial MLE, which is given by $\widehat{p}_{i|j} = \frac{x_{i|j}}{n_j}$. If $\widehat{BC} \geq BC_0$ is observed, then $\widehat{\mathbf{p}} \in \mathcal{S}_0$ which results in

$\sup_{\Theta_0} L(\mathbf{p} \mid \mathbf{x}) = \sup_{\Theta} L(\mathbf{p} \mid \mathbf{x})$. Therefore,

$$\lambda(\mathbf{x}) = \begin{cases} 1 & \text{if } \widehat{BC} \geq BC_0 \\ \frac{sup_{S_0} L(\mathbf{p}|\mathbf{x})}{L(\widehat{\mathbf{p}}|\mathbf{x})} & \text{if } \widehat{BC} < BC_0 \end{cases} \tag{6.10}$$

The degrees of freedom for the test (ν) is the difference of the number of free parameters in the unrestricted parameter space and the restricted parameter space, which is 1. The corresponding p-value for this large sample hypothesis test is

$$p(\mathbf{x}) = \begin{cases} 1 & \text{if } \widehat{BC} \geq BC_0 \\ Pr(\chi_1^2 \geq -2 \log \lambda(x)) & \text{if } \widehat{BC} < BC_0 \end{cases} \tag{6.11}$$

For an observed $\widehat{BC}$ and a fixed BC_0 , the hypothesis is tested by calculating the p-value in Equation 6.11 and comparing this value to the chosen significance level, α . If $p(\mathbf{x})$ is less than α , the null hypothesis is rejected.

6.3.2 One-sided Hypothesis Test on the Difference of Two Bayes Cost Values.

To test the hypothesis

$$H_0 : \eta \leq \eta_0 \text{ vs. } H_1 : \eta > \eta_0 \tag{5.3}$$

using a LRT, the equations from Section 6.2.2 are used, where

$$\eta = \sum_{i=1, i\neq j}^{k} \sum_{j=1}^{k} c_{i|j,A} p_j p_{i|j,A} - \sum_{i=1, i\neq j}^{k} \sum_{j=1}^{k} c_{i|j,B} p_j p_{i|j,B} \tag{6.4}$$

and

$$Y = \widehat{\eta} = \sum_{i=1, i\neq j}^{k} \sum_{j=1}^{k} c_{i|j,A} p_j \frac{X_{i|j,A}}{n_{j,A}} - \sum_{i=1, i\neq j}^{k} \sum_{j=1}^{k} c_{i|j,B} p_j \frac{X_{i|j,B}}{n_{j,B}} \tag{6.5}$$

$\mathcal{A}_{A,B}$, $\mathbf{p_A}$, and $\mathbf{p_B}$ are defined as they were in Section 6.2.2. Recall, the hypothesis to be tested may be written as a restriction on the joint multinomial parameter space.

$$H_0 : (\mathbf{p_A}, \mathbf{p_B}) \in \mathcal{S}_0^2 \text{ vs. } H_1 : (\mathbf{p_A}, \mathbf{p_B}) \in \mathcal{S}_0^{2^C} \tag{6.7}$$

where $\mathcal{S}_0^2 = \{(\mathbf{p_A}, \mathbf{p_B}) : \mathbf{p_A} \in \mathcal{S}, \mathbf{p_B} \in \mathcal{S} \text{ and } \eta \leq \eta_0\}$. The likelihood function for the joint multinomial distribution of both classification systems is

$$L(\mathbf{p_A}, \mathbf{p_B} \mid \mathbf{x_A}, \mathbf{x_A}) \propto \prod_{i=1}^{k} \prod_{j=1}^{k} p_{i|j,A}^{x_{i|j,A}} p_{i|j,B}^{x_{i|j,B}} \tag{6.12}$$

An unrestricted maximization ($\sup_{\Theta} L(\mathbf{p_A}, \mathbf{p_B} \mid \mathbf{x_A}, \mathbf{x_A})$) of this likelihood results in the multinomial MLEs, which are given by $\widehat{p}_{i|j} = \frac{x_{i|j}}{n_j}$. If $\widehat{\eta} \leq \eta_0$ is observed, then $(\widehat{\mathbf{p_A}}, \widehat{\mathbf{p_B}}) \in \mathcal{S}_0^2$ which results in $\sup_{\Theta_0} L(\mathbf{p_A}, \mathbf{p_B} \mid \mathbf{x_A}, \mathbf{x_A}) = \sup_{\Theta} L(\mathbf{p_A}, \mathbf{p_B} \mid \mathbf{x_A}, \mathbf{x_A})$. Therefore,

$$\lambda(\mathbf{x}) = \begin{cases} 1 & \text{if } \widehat{\eta} \leq \eta_0 \\ \frac{sup_{\mathcal{S}_0^2} L(\mathbf{p_A}, \mathbf{p_B} | \mathbf{x_A}, \mathbf{x_A})}{L(\widehat{\mathbf{p_A}}, \widehat{\mathbf{p_B}} | \mathbf{x_A}, \mathbf{x_A})} & \text{if } \widehat{\eta} > \eta_0 \end{cases} \tag{6.13}$$

The degrees of freedom for the test (v) is the difference of the number of free parameters in the unrestricted parameter space and the restricted parameter space, which is 1. The p-value for this hypothesis test is

$$p(\mathbf{x}) = \begin{cases} 1 & \text{if } \widehat{\eta} \leq \eta_0 \\ Pr(\chi_1^2 \geq -2 \log \lambda(x)) & \text{if } \widehat{\eta} > \eta_0 \end{cases} \tag{6.14}$$

For an observed $\widehat{\eta}$ and a fixed η_0 , the hypothesis is tested by calculating the p-value in Equation 6.14 and comparing this value to the chosen significance level, α . For $p(\mathbf{x})$ less than α , the null hypothesis is rejected.

6.4 Simulation Results

A simulation study was conducted to demonstrate the performance of the exact and likelihood ratio hypothesis tests for BC and η . Various scenarios are considered including different sample sizes ($n_j = 5, 10, 20, 30$ for the exact test and $n_j = 10, 50, 100, 250$ for the LRT), differing costs associated with the misclassifications, and classification accuracy (measured by BC_0/η_0 value). All scenarios make no assumptions about the structure of the underlying classification system or feature distributions, and therefore the classification outcomes are simulated with random draws from multinomial distributions. The exact method is appropriate for small sample sizes and the LRT method is appropriate for larger sample sizes which is why they are simulated with different sample size scenarios. However, due to the LRT's good performance at $n_j = 10$, further comparisons between the LRT and exact method are made with small sample sizes using power curves (Section 6.4.1). The performance of the tests is measured by their power and size (Definitions 5 and 6, Section 5.4). Once again, this is accomplished by determining the probability of rejecting the null hypothesis for multiple BC (or η) values.

In Section 6.4.1, the performance of the exact and likelihood ratio one-sided hypothesis tests on a single BC value is evaluated. In Section 6.4.2, the performance of these tests on the difference of two BC values is evaluated. All simulations are run in R assuming a significance level of $\alpha = 0.05$ with 3000 simulation runs [52]. The LRT requires the maximization of the likelihood given the observed data over the null parameter space. This is accomplished by performing a constrained maximization of the multinomial log-likelihood in R using the function *constrOptim* with method "Nelder-Mead" [52].

6.4.1 One-sided Hypothesis Test on a Single Bayes Cost Value.

For consistency, the same BC_0 and cost structures used to demonstrate the performance of the parametric hypothesis tests in Section 5.4.1 are also used in this section. Recall, four BC_0 values are used to demonstrate a range of test performances. Under the assumption of equal costs on all misclassification probabilities, $BC_0 = 0.3, 0.5, 1.0, 1.25$. For the two additional cost structures ($Cost_1 = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}$ and $Cost_2 = \begin{bmatrix} 0 & 2 & 5 \\ 1 & 0 & 3 \\ 1 & 3 & 0 \end{bmatrix}$, $p_j = \frac{1}{3}$) $BC_{0,Cost1} = 0.1, 0.2, 0.35, 0.45$ and $BC_{0,Cost2} = 0.2, 0.4, 0.7, 0.9$. For all simulated BC values, it is assumed the misclassification probabilities are equally distributed among the multinomial misclassification outcomes. The size and power of the exact and likelihood ratio hypothesis tests are presented in Table 6.1 for equal weights, and Tables 6.2 and 6.3 for $Cost_1$ and $Cost_2$, respectively.

It is clear from these results that the exact hypothesis test is an α or smaller sized test (ie. α level test). Also, as expected, the exact hypothesis test is conservative and the power of the test increases as n_j increases (Table 6.1). For $BC_0 = 0.3$ and 0.5 and $n_j = 5$, the test will never reject the null hypothesis. For both of these BC_0 scenarios, the p-values for the tests at $\widehat{BC} = 0$ are 0.21 and 0.06, respectively. Therefore, with $n_j = 5$ these two tests never have enough power to reject the null hypothesis at $\alpha = 0.05$. For $BC_0 = 0.5$, the null hypothesis could be rejected for $\widehat{BC} = 0$ with a significance level greater than 0.06. Similar scenarios with respect to p-values and power result for the exact test for $Cost_1$ and $Cost_2$ (Tables 6.2 and 6.3). Notably, these cost structures result in decreased power for the exact test.

The LRT ($n_j \geq 10$) is also an α level test (Tables 6.1 through 6.3). Like the exact test, the power increases for increasing n_j . There are also scenarios where this test never rejects the null

hypothesis, due to comparable reasons as the exact test (Tables 6.2 and 6.3). Finally, when the costs on the misclassification probabilities are not equal, the LRT generally has higher power than the exact test when considering the same sample size scenario ($n_j = 10$), with some exceptions for small BC_0 values.

To consider the comparison between the exact test and LRT further, power curves were plotted for differing BC_0 values assuming equal costs and small sample size scenarios (Figure 6.1). These plots visually demonstrate the similar performance between both hypothesis test methods. Although the LRT is more powerful than the exact test at $n_j = 5$, the LRT also has size greater than α at this sample size. For larger sample sizes considered with the power curves ($n_j = 20, 30$) the exact test is more powerful than the LRT (see Figure 6.1). Also, it is clear from these power curves that detecting a more accurate classification system (smaller BC_0 value), requires larger sample sizes.

Table 6.1: Power when the misclassifications have equal weights. Detectable difference indicates the difference of the assumed true BC value and BC_0 ($BC > BC_0$). The power at a detectable difference of zero is the estimated size of the hypothesis test.

	Detectable	Exact Hypothesis Test				Likelihood Ratio Test			
	Difference	n_j =5	10	20	30	10	50	100	250
$BC_0 = 0.30$	0 (α)	0.000	0.045	0.017	0.051	0.037	0.031	0.027	0.026
	0.01	0.000	0.049	0.017	0.060	0.045	0.040	0.042	0.051
	0.05	0.000	0.070	0.033	0.120	0.071	0.110	0.168	0.349
	0.10	0.000	0.128	0.088	0.270	0.128	0.317	0.554	0.911
	0.20	0.000	0.360	0.402	0.815	0.367	0.929	0.997	1.000
	0.30	0.000	1.000	1.000	1.000	0.992	1.000	1.000	1.000
$BC_0 = 0.50$	0 (α)	0.000	0.030	0.024	0.023	0.027	0.027	0.025	0.026
	0.01	0.000	0.033	0.025	0.028	0.029	0.034	0.034	0.044
	0.05	0.000	0.057	0.046	0.063	0.041	0.077	0.110	0.228
	0.10	0.000	0.082	0.089	0.143	0.068	0.192	0.339	0.722
	0.20	0.000	0.183	0.260	0.448	0.169	0.662	0.920	1.000
	0.30	0.000	0.395	0.616	0.855	0.389	0.975	1.000	1.000
	0.40	0.000	0.729	0.949	0.997	0.727	1.000	1.000	1.000
	0.50	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$BC_0 = 1.00$	0 (α)	0.023	0.038	0.037	0.044	0.036	0.027	0.026	0.023
	0.01	0.021	0.038	0.041	0.050	0.039	0.031	0.034	0.039
	0.05	0.028	0.051	0.063	0.088	0.054	0.062	0.089	0.155
	0.10	0.040	0.080	0.099	0.152	0.080	0.132	0.235	0.500
	0.20	0.066	0.156	0.223	0.368	0.153	0.418	0.720	0.980
	0.30	0.111	0.260	0.445	0.658	0.263	0.773	0.973	1.000
	0.40	0.170	0.412	0.688	0.887	0.429	0.960	0.999	1.000
	0.50	0.256	0.610	0.882	0.978	0.617	0.997	1.000	1.000
$BC_0 = 1.25$	0 (α)	0.017	0.028	0.046	0.043	0.028	0.022	0.025	0.020
	0.01	0.022	0.033	0.051	0.049	0.031	0.026	0.032	0.032
	0.05	0.025	0.041	0.076	0.083	0.044	0.056	0.084	0.150
	0.10	0.037	0.066	0.118	0.142	0.065	0.119	0.211	0.462
	0.20	0.060	0.118	0.246	0.324	0.129	0.375	0.661	0.969
	0.30	0.106	0.210	0.444	0.599	0.220	0.705	0.947	1.000
	0.40	0.160	0.343	0.672	0.819	0.364	0.923	0.997	1.000
	0.50	0.235	0.506	0.842	0.947	0.520	0.992	1.000	1.000

Table 6.2: Power when the misclassifications have a cost structure given by $Cost_1$. Detectable difference indicates the difference of the assumed true BC value and BC_0 ($BC < BC_0$). The power at a detectable difference of zero is the estimated size of the hypothesis test.

	Detectable	Exact Hypothesis Test			Likelihood Ratio Test			
	Difference	n_j =5	10	20	10	50	100	250
$BC_0 = 0.10$	0 (α)	0.000	0.000	0.039	0.000	0.037	0.028	0.023
	0.01	0.000	0.000	0.057	0.000	0.070	0.073	0.118
	0.05	0.000	0.000	0.248	0.000	0.451	0.747	0.990
	0.10	0.000	0.000	0.984	0.000	1.000	1.000	1.000
$BC_0 = 0.20$	0 (α)	0.000	0.043	0.020	0.034	0.027	0.027	0.023
	0.01	0.000	0.047	0.023	0.044	0.045	0.055	0.082
	0.05	0.000	0.100	0.079	0.094	0.240	0.415	0.802
	0.10	0.000	0.251	0.332	0.262	0.772	0.972	1.000
	0.15	0.000	0.571	0.784	0.565	0.996	1.000	1.000
	0.20	0.000	1.000	1.000	1.000	1.000	1.000	1.000
$BC_0 = 0.35$	0 (α)	0.010	0.013	0.024	0.045	0.025	0.026	0.032
	0.01	0.012	0.019	0.024	0.051	0.042	0.045	0.070
	0.05	0.022	0.028	0.070	0.095	0.169	0.278	0.595
	0.10	0.045	0.083	0.210	0.183	0.518	0.817	0.995
	0.15	0.087	0.190	0.466	0.331	0.883	0.995	1.000
	0.20	0.169	0.387	0.761	0.539	0.995	1.000	1.000
	0.25	0.306	0.641	0.956	0.773	1.000	1.000	1.000
	0.30	0.573	0.887	1.000	0.950	1.000	1.000	1.000
$BC_0 = 0.45$	0 (α)	0.011	0.023	0.020	0.045	0.026	0.023	0.026
	0.01	0.012	0.028	0.025	0.055	0.040	0.043	0.056
	0.05	0.029	0.056	0.066	0.096	0.149	0.241	0.529
	0.10	0.051	0.122	0.177	0.169	0.446	0.739	0.983
	0.15	0.082	0.224	0.381	0.291	0.804	0.979	1.000
	0.20	0.145	0.389	0.637	0.445	0.974	0.998	1.000
	0.25	0.240	0.604	0.865	0.636	0.998	1.000	1.000
	0.30	0.374	0.800	0.978	0.813	1.000	1.000	1.000

Table 6.3: Power when the misclassifications have a cost structure given by $Cost_2$. Detectable difference indicates the difference of the assumed true BC value and BC_0 ($BC < BC_0$). The power at a detectable difference of zero is the estimated size of the hypothesis test.

	Detectable	Exact Hypothesis Test			Likelihood Ratio Test			
	Difference	n_j =5	10	20	10	50	100	250
$BC_0 = 0.20$	0 (α)	0.000	0.000	0.038	0.000	0.040	0.031	0.030
	0.01	0.000	0.000	0.043	0.000	0.051	0.045	0.057
	0.05	0.000	0.000	0.094	0.000	0.156	0.215	0.451
	0.10	0.000	0.000	0.218	0.000	0.427	0.697	0.980
	0.15	0.000	0.000	0.500	0.000	0.829	0.987	1.000
	0.20	0.000	0.000	1.000	0.000	1.000	1.000	1.000
$BC_0 = 0.40$	0 (α)	0.000	0.000	0.024	0.026	0.030	0.029	0.027
	0.01	0.000	0.000	0.030	0.029	0.041	0.039	0.048
	0.05	0.000	0.000	0.058	0.046	0.101	0.146	0.285
	0.10	0.000	0.000	0.113	0.086	0.252	0.436	0.815
	0.15	0.000	0.000	0.211	0.157	0.512	0.804	0.994
	0.20	0.000	0.000	0.364	0.268	0.799	0.979	1.000
	0.25	0.000	0.000	0.576	0.419	0.963	1.000	1.000
	0.30	0.000	0.000	0.805	0.646	1.000	1.000	1.000
$BC_0 = 0.70$	0 (α)	0.000	0.019	0.024	0.046	0.029	0.030	0.024
	0.01	0.000	0.023	0.030	0.048	0.035	0.037	0.036
	0.05	0.000	0.032	0.052	0.065	0.077	0.102	0.183
	0.10	0.000	0.050	0.087	0.088	0.165	0.272	0.601
	0.15	0.000	0.075	0.158	0.131	0.308	0.534	0.898
	0.20	0.000	0.110	0.226	0.188	0.497	0.789	0.991
	0.25	0.000	0.154	0.354	0.241	0.695	0.944	1.000
	0.30	0.000	0.219	0.490	0.319	0.855	0.989	1.000
$BC_0 = 0.90$	0 (α)	0.011	0.018	0.020	0.043	0.028	0.026	0.023
	0.01	0.010	0.022	0.023	0.046	0.033	0.035	0.037
	0.05	0.018	0.032	0.049	0.060	0.072	0.093	0.162
	0.10	0.024	0.048	0.081	0.087	0.147	0.242	0.525
	0.15	0.032	0.074	0.125	0.129	0.270	0.467	0.851
	0.20	0.046	0.117	0.190	0.168	0.435	0.711	0.980
	0.25	0.056	0.156	0.288	0.213	0.616	0.890	1.000
	0.30	0.070	0.213	0.417	0.265	0.771	0.974	1.000

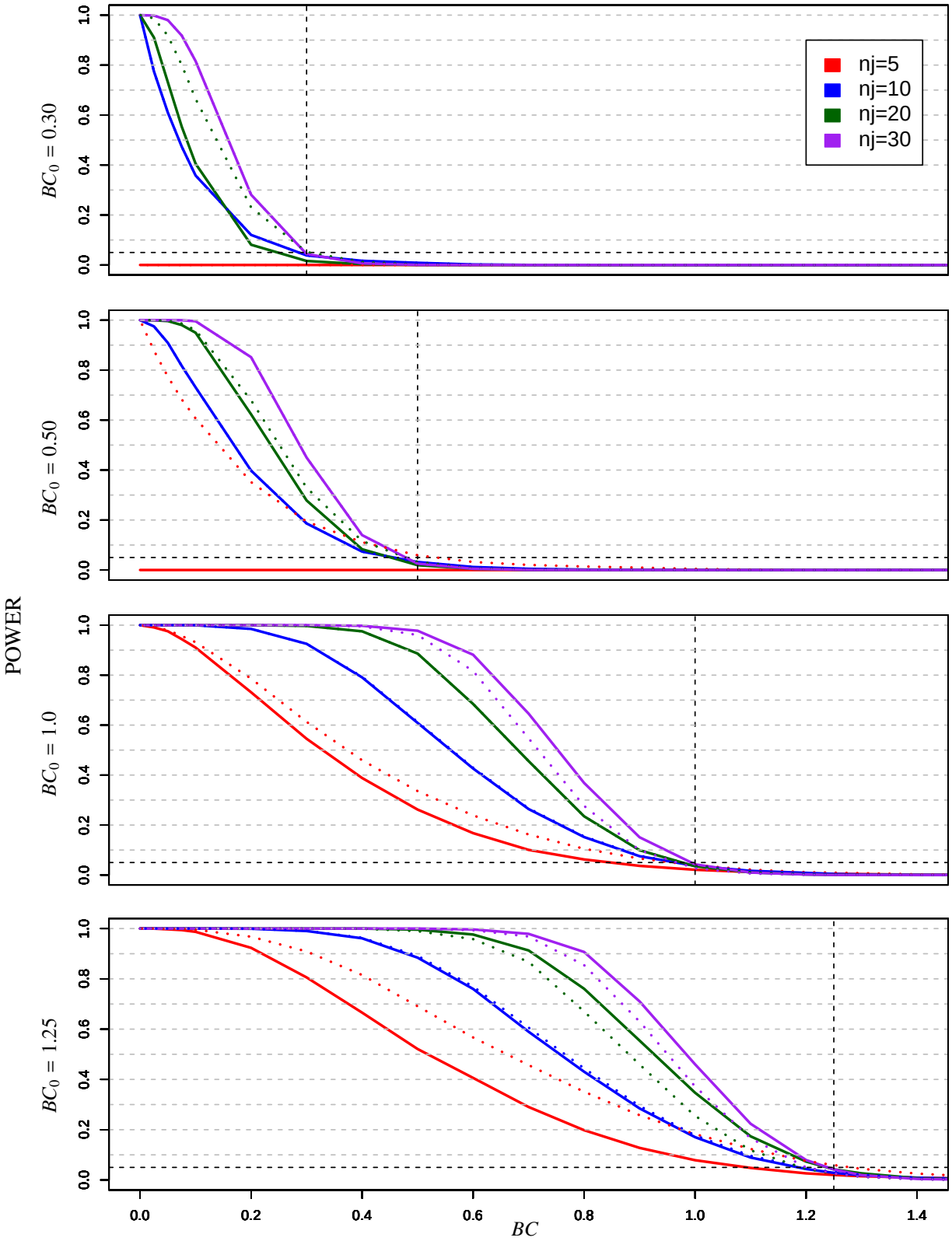

Figure 6.1: Power curves for Exact (solid line) and Likelihood Ratio (dashed line) hypothesis tests for $n_j = 5$ (red), $n_j = 10$ (blue), $n_j = 20$ (green), and $n_j = 30$ (purple) at different BC_0 values.

6.4.2 *One-sided Hypothesis Test on the Difference of Two Bayes Cost Values.*

For testing the difference of two independent classification systems, $\eta_0 = 0$ is used. To consider different detectable differences for the test, BC_A is fixed at 0.8 and BC_B is varied ($BC_B = (0.3, \ldots, 0.8)$) to simulate the desired η values. Multinomial random variables are generated assuming the misclassification probabilities are evenly distributed among the classes for all BC values.

For the exact hypothesis test, the sample space for two independent, three-class classification systems ($\mathcal{A}_{A,B}$, to consider BC_A and BC_B simultaneously) becomes very large. Due to this large sample space, the computational time is also large. Therefore, the test is run for small sample sizes only and assuming all $c_{i|j}p_j = 1$, for $i \neq j$ (allowing for binomial distributions to be used instead of multinomial distributions, in order to reduce the sample space). The results are presented in Table 6.4. Both the exact and LRT hypothesis tests perform similarly with respect to power and sample size, although for $n_j = 10$ the exact test is more powerful than the LRT. Also, both tests have size $\leq \alpha$.

Table 6.4: Power for multinomial distributed classes with equal weights for testing $\eta \leq 0$. Detectable difference indicates the difference of the assumed true value of $BC_A - BC_B$ ($\eta \geq 0$). The power at a detectable difference of zero is the estimated size of the hypothesis test.

	Detectable	Exact Test		Likelihood Ratio Test			
	Difference	n_j =5	10	10	50	100	250
$\eta_0 = 0$	0 (α)	0.045	0.038	0.042	0.031	0.030	0.029
	0.01	0.052	0.042	0.044	0.039	0.039	0.033
	0.05	0.072	0.078	0.051	0.063	0.078	0.102
	0.10	0.112	0.141	0.070	0.114	0.169	0.315
	0.20	0.197	0.298	0.125	0.301	0.496	0.854
	0.30	0.326	0.540	0.208	0.585	0.866	0.997
	0.40	0.507	0.752	0.332	0.854	0.989	1.000
	0.50	0.687	0.910	0.493	0.977	1.000	1.000

6.5 Summary

Two nonparametric methods for testing hypotheses on BC were derived, an exact test for small sample sizes and a LRT based on large sample theory. An interesting result from the simulation is the similar performance of the exact and LRT hypothesis tests, especially in the hypothesis test on a single BC value. Although the LRT is an approximate method, it performs similar to the exact test with respect to power, even for the $n_j = 10$ small sample size. Due to the discrete sample space of $\widehat{BC}$, although the p-values found with the LRT test are approximate, they are accurate enough to make the same decision as the exact test for some observed values of BC. This is demonstrated for an example in Table 6.5, for testing different BC_0 values for a three-class classification system with $n_j = 10$ and $\widehat{BC} = 0.1$. In this example, although the LRT p-values are not the same as the exact p-values, they result in the same decision (with respect to rejecting or failing to reject the null hypothesis) for $\alpha = 0.05$. Consequently, the two methods at times have similar performance with respect to size and power.

Table 6.5: P-values for exact and likelihood ratio tests for a three-class scenario for testing a single BC_0 value with $n_j = 10$ and $\widehat{BC} = 0.1$

BC_0	Exact p-value	LRT p-value
0.3	0.184	0.127
0.5	0.029	0.011
1	8.34E-05	1.52E-05
1.25	2.89E-08	3.14E-07

Another result of interest is that when the misclassification weights are unequal, the likelihood ratio test generally has slightly higher power than the exact test (although notably this comparison is only made for $n_j = 10$). The exact hypothesis test was implemented to calculate a p-value by searching the null probability space, incremented by probabilities of 0.05. Therefore, a better search method for finding these exact p-values may result in more precise (less conservative) values which could increase the power of this test.

The methods developed in this section provide flexible hypothesis tests which may be used for testing the performance of a single classification system or for comparing performance between classification systems. These hypothesis tests may be implemented despite differing classification structures or nonparametric scenarios. The exact hypothesis tests perform well, but are computationally difficult for increasing sample size (especially for tests on η). The LRTs therefore provide an approximate alternative to the exact test that is easier to implement computationally, especially for larger n_j .

VII. Applications

7.1 Classifying Breast Cancer

The methods proposed in Chapter 3 are used to distinguish classes of the Breast Tissue data set from the UCI Machine Learning Repository [4]. This data set consists of 106 observations of nine continuous features derived from electrical impedance spectroscopy truncated spectrum of breast tissue, which have been shown to discriminate breast tissue into six categories: Carcinoma (CAR, n=21), Fibro-adenoma (FAD, n=15), Mastopathy (MAS, n=18), Glandular (GLA, n=16), Connective (CON, n=14), and Adipose (ADI, n=22) [61]. By grouping the classes GLA, FAD, and MAS together (denoted FAD+MAS+GLA) this becomes a four-class classification problem. These three classes are grouped together because their discrimination is not considered important and they cannot be discriminated using the available features [4, 61]. In [61], linear discriminant analysis was used to distinguish between various subgroups of classes and it was determined that the low frequency limit (I_0), area under the spectrum normalized by impedance distance between spectral ends ($AREA_{D_A}$), and the maximum of the spectrum (IP_{max}) were the best features for discriminating between freshly excised breast tissue. However, it was also suggested that the length of the spectral curve feature (P), may be able to simultaneously discriminate between the four derived classes of interest [61]. This four-class diagnostic scenario is addressed using the derived parametric methods, considering these four features as potential class discriminators (I_0, $AREA_{D_A}$, IP_{max}, and P).

Mean, standard deviation, median, and range of the four features for each class are presented in Table 7.1. P appears to have small overlap between all groups when compared to the other features, indicating it may perform well as a classifier. IP_{Max} and I_0 have significant overlap between the CAR group and at least one other feature (CON for IP_{Max} and FAD+MAS+GLA for I_0). $AREA_{D_A}$ has substantial overlap between all classes. The methods developed in Chapter 3 require normality of the feature to be used for classification, however the mean and median data indicates that some of the features may be skewed. The Shapiro-Wilk test is used to test this assumption and performs well compared to other goodness of fit tests [32]. The assumption of normality is met for IP_{max}

only, so a Box-Cox transformation is used to transform the other three features to normality where

$$\text{Feature}_{\text{transformed}} = \frac{\text{Feature}^{\lambda} - 1}{\lambda} \tag{7.1}$$

This results in $\lambda = 0.09$ for $AREA_{D_A}$ and $\lambda = -0.31$ for both I_0 and P, found using the *powerTransform* function in the *car* package in R [24, 52, 60]. After the transformation, all classes pass the test for normality except for connective tissue with a p-value of .014 and .047 in I_0 and P, respectively. As was demonstrated in Chapter 3, these slight deviations from normality are not expected to have a large negative impact on the CI around BC, however the CIs around the optimal thresholds may not perform well.

Prevalences are adjusted to account for the FAD+MAS+GLA class being the combination of three classes, resulting in prevalences of: $p_{FAD+MAS+GLA} = \frac{1}{2}$ and $p_{CAR} = p_{CON} = p_{ADI} = \frac{1}{6}$. All four features ($I_0$, $AREA_{D_A}$, IP_{max}, and P) are considered separately as potential features to discriminate between the four classes (with equal cost given to all misclassification rates). For each feature, BC_4 and its 95% CI is determined using Equation 2.16 and the GCI presented in Section 3.3.3 where

$$P_{1|j} = \Phi\left(\frac{\theta_1 - \mu_j}{\sigma_j}\right) \tag{7.2}$$

$$P_{2|j} = \Phi\left(\frac{\theta_2 - \mu_j}{\sigma_j}\right) - \Phi\left(\frac{\theta_1 - \mu_j}{\sigma_j}\right) \tag{7.3}$$

$$P_{3|j} = \Phi\left(\frac{\theta_3 - \mu_j}{\sigma_j}\right) - \Phi\left(\frac{\theta_2 - \mu_j}{\sigma_j}\right) \tag{7.4}$$

$$P_{4|j} = \Phi\left(\frac{\mu_j - \theta_3}{\sigma_j}\right) \tag{7.5}$$

and Φ is the standard normal CDF [52]. The GCIs are chosen for this application over the delta method CIs due to the sample sizes in each class.

Because the CAR class may be considered the most important to detect, a second cost structure is assumed which gives greater cost for misclassifying a CAR subject as any of the other classes and also a higher cost on the class specific misclassification of any subjects from the other three classes as CAR. This results in a cost structure where $Cost = \begin{bmatrix} 0 & 9 & 4 & 4 \\ 6 & 0 & 6 & 6 \\ 4 & 9 & 0 & 4 \\ 4 & 9 & 4 & 0 \end{bmatrix}$, assuming an ordering of the class means where $\mu_{FAD+MAS+GLA} < \mu_{CAR} < \mu_{CON} < \mu_{ADI}$ (the cost structure is adjusted appropriately for features with a different ordering). Once again the BC_4 value and associated 95% GCI for all

four features are determined. The BC_4 values and 95% CI for each feature and cost structure are given in Table 7.1.

Table 7.1: Descriptive statistics for features (broken into four classes: FAD+MAS+GLA, CAR, CON, ADI) to classify breast tissue and each features' BC_4 values with 95% generalized confidence intervals.

Feature		Mean	Standard Deviation	Median	Range
P	FAD+MAS+GLA	283.38	106.30	252.48	[124.98, 553.38]
	CAR	479.97	93.19	477.55	[329.09, 656.77]
	CON	1065.00	356.07	1121.19	[528.70, 1524.61]
	ADI	2138.75	386.51	2068.05	[1475.37, 2896.52]
	BC_4 equal costs		0.65 (0.49, 0.91)		
	BC_4 unequal costs		1.02 (0.75, 1.46)		
IP_{Max}	FAD+MAS+GLA	27.20	10.22	26.86	[7.97, 49.33]
	CAR	64.53	18.85	69.39	[35.60, 96.56]
	CON	72.96	34.45	70.10	[23.98, 143.09]
	ADI	194.60	106.56	164.63	[51.85, 436.10]
	BC_4 equal costs		0.89 (0.73, 1.16)		
	BC_4 unequal costs		1.32 (1.08, 1.74)		
I_0	FAD+MAS+GLA	259.73	104.22	245	[103.00, 544.65]
	CAR	394.23	87.04	389.87	[269.50, 551.88]
	CON	1212.86	386.47	1328.17	[649.37, 1724.09]
	ADI	2052.05	342.49	1974.56	[1600.00, 2800.00]
	BC_4 equal costs		0.77 (0.58, 1.04)		
	BC_4 unequal costs		1.21 (0.90, 1.63)		
$AREA_{D_A}$	FAD+MAS+GLA	10.25	6.60	9.19	[2.76, 33.60]
	CAR	32.05	9.28	31.30	[15.94, 44.90]
	CON	14.00	10.77	14.77	[1.60, 43.39]
	ADI	50.78	33.93	44.59	[14.64, 164.07]
	BC_4 equal costs		1.31 (1.16, 1.52)		
	BC_4 unequal costs		2.14 (1.99, 2.62)		

Using the BC_4 values and their 95% CIs, discriminatory ability of each feature is determined (equal or unequal costs). All features perform better than chance. It is clear that P and IP_{max} are performing better than $AREA_{D_A}$ for equal and unequal costs since the CIs around BC_4 for $AREA_{D_A}$ are higher than the other two. Under the carcinoma weighted cost structure the CIs for P, IP_{Max},

and I_0 overlap and therefore these features may be considered equally good. However, for both cost structures considered, P has the lowest estimate for BC_4 and it also has the lowest upper bound on the 95% CI, indicating the lowest maximum potential BC_4 value.

Choosing P to discriminate between all four classes with equal costs (with $\mu_{FAD+MAS+GLA} < \mu_{CAR} < \mu_{CON} < \mu_{ADI}$) , the optimal thresholds ($\theta_1^* < \theta_2^* < \theta_3^*$) and their 95% GCIs are θ_1^*= 402.21 (375.13, 441.14), θ_2^*= 643.20 (587.21, 717.27), and θ_3^*=1540.50 (1387.497, 1665.80). The contingency table resulting from applying this classifier at its optimal point to the data is presented in Table 7.2. Choosing P to discriminate between all four classes with a higher cost on the misclassification of carcinoma, the optimal thresholds and their 95% GCIs are θ_1^* = 380.53 (353.07, 409.69), θ_2^* = 662.83 (596.61, 740.75), and θ_3^* = 1540.50 (1397.89, 1675.78). The contingency table resulting from applying this classifier to the data at its optimal point is also presented in Table 7.2. The two different cost structures result in different estimates for θ_1^* and θ_2^* , but not for θ_3^* , demonstrating the impact differing cost structures may have on determining the optimal thresholds.

Table 7.2: Contingency tables for classifying breast tissue using length of spectral curve (P).

	Predicted Class	True Class			
		FAD+MAS+GLA	CAR	CON	ADI
Equal Costs	FAD+MAS+GLA	0.90	0.24	0.00	0.00
	CAR	0.10	0.71	0.21	0.00
	CON	0.00	0.05	0.79	0.05
	ADI	0.00	0.00	0.00	0.95
		FAD+MAS+GLA	CAR	CON	ADI
Unequal Costs	FAD+MAS+GLA	0.80	0.14	0.00	0.00
	CAR	0.20	0.86	0.29	0.00
	CON	0.00	0.00	0.71	0.05
	ADI	0.00	0.00	0.00	0.95

Using the thresholds which result from the cost structure which weights the misclassification of carcinoma higher, the correct classification rate for carcinoma increases from 71% to 86%. This results in 14% of CAR subjects being misclassified as FAD+MAS+GLA (also an abnormal state). None of the carcinoma cases are being classified as either of the two normal classes (CON and

ADI) when the weighted cost structure is used. In [61], linear discriminant analysis was used for the classification of subgroups of the six classes. Using this method, more than one feature may be considered at a time for discrimination. When discriminating only between two classes, CAR and FAD+MAS+GLA, they found two features ($AREA_{D_A}$ and IP_{max}) resulted in the best classifier. Using this linear discrimination they had approximately the same correct classification rate for CAR (86.36%) as we observed. However, our diagnostic tests are simpler (depend on one feature using simple cut-offs between classes) and simultaneously classifies between all four classes. If distinctions between only CAR and FAD+MAS+GLA were of interest, higher correct classification rates may potentially be achieved using other features. Using linear discriminant analysis, the false negative rate may be altered by adjusting boundaries for a single class of interest, however costs for all decisions can not be accounted for *a priori*. Finally, the resulting classification rates for the connective tissue group are the worst, which may be a result of this group's departure from normality.

The CIs around BC reflect the uncertainty in each feature's ability to classify due to the variation of the data. Notably, as observed from the simulation results, the CI on BC is more robust than the CIs on the optimal thresholds for transformed data in the Box-Cox family (as in this application). Here, constructing a CI on BC allows the researcher to decide on the best feature (or test). In this study, P was found to be the best single feature for classifying breast tissue. Further study may be conducted in order to verify the optimal thresholds to implement this feature in practice for diagnosis.

7.2 Classifying Chronic Allograft Nephropathy

After kidney transplant (KT), chronic allograft nephropathy (CAN) is one of the prevalent factors leading to renal transplant failure, yet its progression is still not well understood. Biopsy is a means of determining if a patient has CAN, however it is of interest to determine methods for detecting progression towards CAN after KT which are less invasive. Due to the inflammatory response generated by tissue damage associated with CAN, it has been suggested that proinflammatory cytokine markers, such as the transforming growth factor-β1 may provide an early indication of potential allograft loss [48]. Mas et. al. conducted a study to evaluate gene panel mRNAs in urine samples for their usefulness as a non-invasive tool for evaluating graft function [37]. This study suggested that the biomarkers transforming growth factor-β1 (TGF-β1), angiotensinogen (AGT), and epidermal growth factor receptor (EGFR) (all measurable mRNA levels in urine) could be useful as early predictors of allograft function [37]. There were 32 normal kidney function patients (NKF) , 18 normal kidney function with proturina patients (NKF+, a progression towards CAN), and 14 CAN patients six months post transplant examined in their study. Descriptive statistics of the three biomarkers within each diagnostic state are presented in Table 7.3 with a more detailed description of all the markers originally considered found in [37].

Table 7.3: Descriptive statistics of three features (broken into three classes: NKF, NKF+, CAN) to classify kidney function.

Feature	Class	Mean	Standard Deviation	Median	Range
AGT	NKF	15.47	16.02	8.02	[1,64]
	NKF+	4.76	6.30	2.90	[0.11,24.25]
	CAN	4.63	3.44	4.15	[0.05,9.85]
$TGF-\beta1$	NKF	1.56	1.22	1.37	[0.13,6.06]
	NKF+	32.75	128.85	1.04	[0.33,548.75]
	NKF+†	2.39	4.58	0.93	[0.33,19.70]
	CAN	5.31	5.06	3.26	[1.23,19.70]
EGFR	NKF	15.41	15.34	9.71	[1,64]
	NKF+	7.12	12.51	4.01	[0.11,51.98]
	CAN	4.23	3.27	3.65	[0.05,9.85]

†These values exclude the extreme observation where TGF-β1 = 548.74.

Potential multi-class classifiers were evaluated in [57] using volume under the surface (VUS) of the ROC manifold. The highest VUS (best classification performance) resulted from a classifier which simultaneously utilized both the AGT and TGF-β1 biomarkers, splitting the two dimensional parameter space into regions for classification using arrays. However, the mathematical complexity of this classifier makes it hard to implement.

Instead, a simplified version of the classifier in [57] with practical rules using thresholds for the observed values of AGT and TGF-β1 may be used. Further, comparisons between different classifiers utilizing such rules, with varying levels of complexity are made. First, Classifier 1 is a simpler classifier, utilizing single threshold values on the two biomarkers for TGF-β1 and AGT, respectively ($\boldsymbol{\theta} = (\theta_1, \theta_2)$):

Classifier 1:

Assign patient i to

class 3 (CAN) if $x_{TGF-\beta,i} > \theta_1$

class 2 (NKF+) if $x_{TGF-\beta,i} \leq \theta_1$ and $x_{AGT,i} < \theta_2$

or class 1 (NKF) otherwise.

This classifier is plotted in Figure 7.1 (top) using the optimal threshold values which were found to minimize the empirically estimated BC using a simple grid search. These threshold values associated with the minimum BC (equal costs and prevalences are assumed for all misclassification outcomes) are $\boldsymbol{\theta} = (2.55, 3.65)$. This classifier is represented with vertical and horizontal lines and has the advantage of only requiring two threshold values. For example, a subject whose TGF-β1 is 2.4 and an has AGT of 3.1 would be classified with NKF+ and a subject whose TGF-β1 is greater than 2.55, regardless of their AGT value, would be classified with CAN. Classifier 1 correctly classified 26 of 32 patients as NKF, 9 of 18 patients as NKF+, and 11 of 14 patients as CAN and has a corresponding $\widehat{BC} = 0.90$ (see Table 7.4 for the full contingency table of classification outcomes).

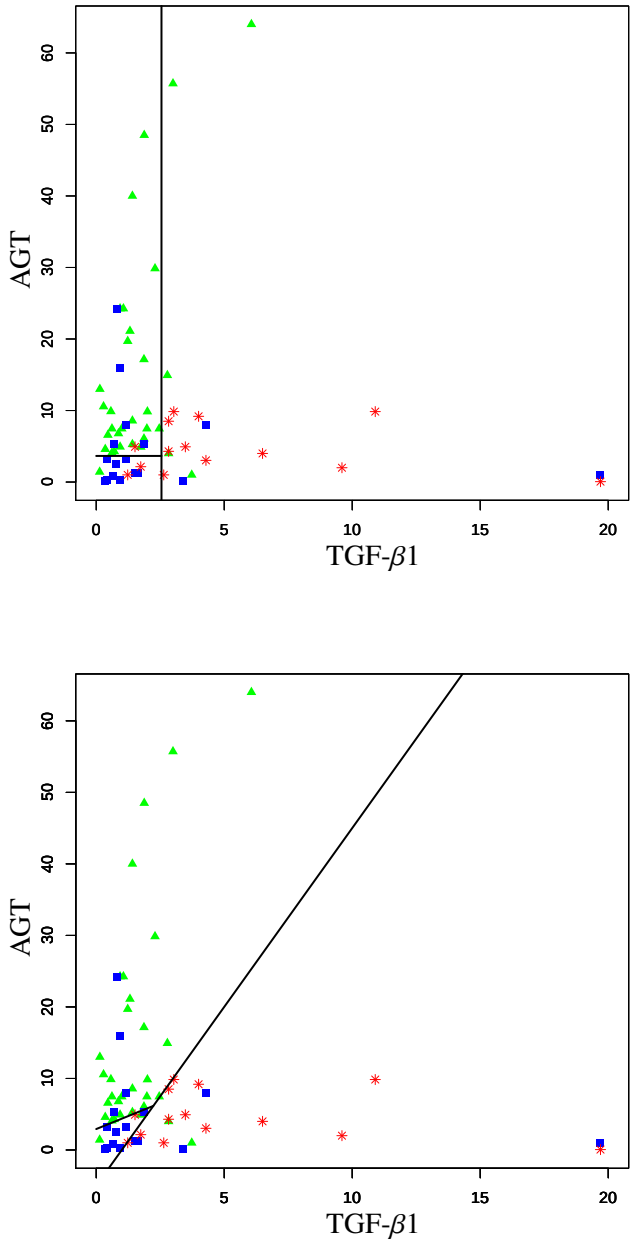

Figure 7.1: Plot of AGT vs. TGF-β1 with three-class classification systems (Top: Classifier 1, Bottom: Classifier 2) for classifying patients as NKF (▲), NKF+ (■), or CAN (*). These plots exclude the extreme observation in TGF-β1, where (TGF-β1, AGT)=(548.74, 4.59), however this point is included in the classification.

A more complex variant of Classifier 1 is also proposed that allows for the horizontal and vertical lines to have slope. This classifier, Classifier 2, considers non-rectangular regions in the AGT and TGF-β1 plane and requires four thresholds ($\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)$):

Classifier 2:

Assign patient i to

class 1 (NKF) if $x_{AGT,i} > \left[\theta_4 \times x_{TGF-\beta1,i} - \theta_4\theta_3\right]$ and $x_{AGT,i} > \left[\theta_2 \times x_{TGF-\beta1,i} + \theta_1\right]$

class 3 (CAN) if $x_{AGT,i} \leq \left[\theta_2 \times x_{TGF-\beta1,i} + \theta_1\right]$

or class 2 (NKF+) otherwise.

This classifier is plotted in Figure 7.1 (bottom) using the four optimal threshold values associated with the minimum $\widehat{BC}$, $\boldsymbol{\theta} = (2.925, 1.45, 1.0, 5.0)$. Classifier 2 correctly classified 28 of 32 patients as NKF, 8 of 18 patients of NKF+, and 13 of 14 patients as CAN with a corresponding $\widehat{BC} = 0.75$ (see Table 7.4 for the full contingency table of classification outcomes). Based on these point estimates of BC, Classifier 2 is performing better than Classifier 1, demonstrating the potential utility of non-rectangular regions in this instance.

Table 7.4: Contingency tables for classifying subjects into three groups with respect to chronic allograft nephropathy.

	Predicted Class	True Class		
		NKF	NKF+	CAN
Classifier 1	NKF	0.81	0.28	0.07
$\widehat{BC} = 0.90$	NKF+	0.03	0.50	0.14
	CAN	0.16	0.22	0.79
		NKF	NKF+	CAN
Classifier 2	NKF	0.88	0.22	0.00
$\widehat{BC} = 0.76$	NKF+	0.06	0.44	0.07
	CAN	0.06	0.33	0.93

This data consists of small sample sizes of the classes, non-normality of the biomarkers in each class (which do not transform to normality), and the requirement to use two biomarkers simultaneously in order to make the desired classifications. Therefore, the proposed fiducial interval from Chapter 4 can be used to construct a CI around the optimal BC for both classifiers.

Using the fiducial interval, a 95% CI for Classifier 1 is $BC \in [0.56, 1.29]$ and for Classifier 2 is $BC \in [0.44, 1.13]$. Both CIs demonstrate that these classifiers are performing better than chance because they do not span $BC = 1.5$. Although Classifier 2 reflects better classification for the CAN diagnostic state (13 instead of 11 patients correctly classified), the overlap of these two CIs indicates that Classifier 2 may not perform better than Classifier 1 across all diagnostic states.

A nonparametric hypothesis test may be conducted to formally test whether the more complex classifier (Classifier 2) is performing better than the simpler classifier (Classifier 1). This was accomplished with the LRT developed in Section 6.3.2 for testing hypotheses on η . Based on the simulation results in Section 6.4.2, the LRT is appropriate for this application because for sample sizes of $n_j = 10$ or more the LRT maintained a size less than α. For this application,

$$\eta = BC_{Classifier1} - BC_{Classifier2} \tag{7.6}$$

and the hypothesis being tested is

$$H_0 : \eta \leq \eta_0 \text{ vs. } H_1 : \eta > \eta_0 \tag{5.3}$$

Using the LRT, the p-value for this test is 0.51 ($\widehat{\eta} = 0.15$). The exact hypothesis test was shown with simulations in Section 6.4.2 to have higher power than the LRT for tests on η . However, although applying the exact test here might result in a slightly smaller p-value, the difference in p-values would not be enough to change the decision of the test at a significance level of 0.05. Therefore, the null hypothesis is not rejected and there is not enough evidence to conclude the more complex classifier is performing better than the simpler classifier.

This application demonstrates the use of nonparametric inference methods on BC for a classifier using thresholds for a pair of biomarkers. Future work on associating the inflammatory response with diagnostic states leading to CAN, may utilize these methods to make comparisons between combinations of alternate classifiers (e.g. random forests) and biomarkers to determine that which best aids diagnosis of allograft function post transplant. This demonstrates an important use of flexible inference methods for BC.

VIII. Conclusions

Performance of classification systems at their optimal point is of great importance for classification methods. The commonly employed Youden index allows for summarizing a classification system's performance at its optimal thresholds, as the sum of correct classification rates. Bayes Cost, which minimizes misclassification rates instead, has been shown to be a more flexible metric for characterizing performance of a classification system due to its ability to allow for any costs and prevalence to be placed on all class specific misclassifications. In fact, due to the flexibility of BC, the methods developed in this dissertation may also be used for inference on J.

Although estimating BC and the optimal thresholds is of interest, quantifying the uncertainty in a classification system's performance is also of great practical use, especially if the classification system is not already determined, or if new or varying tests require comparison. Therefore, this work has developed new CI and hypothesis test methods for BC under parametric and nonparametric frameworks. CIs for $k \geq 3$ classes were limited in the literature, and previous to this work, hypothesis tests had not been developed. Under parametric scenarios, the generalized inference methods were shown with simulation to outperform the inference methods which utilized the delta method. For nonparametric settings, exact inference methods were derived which were developed with the fiducial argument. These methods may require large computational time, and therefore a likelihood ratio test was also developed which may be used as an approximate alternative to the exact hypothesis test when sample sizes are large enough. The methods which have been proposed are possible for any finite number of outcome classes.

BC can incorporate any cost structure on the correct and incorrect classification rates. However, it is possible to pick cost structures that would result in no optimal solution for the classification system [65]. Therefore, costs should be chosen with realistic concerns in mind. If costs reflect truth and no solution exists for the classification system, then the costs must be adjusted if possible, or more ideally, a better system found which can allow for the necessary cost structure.

Future work may consider more efficient methods for calculating the exact fiducial interval bounds as well as computing exact p-values, therefore conserving computational time and making

the implementation of the exact methods easier. Also, the GCI performed well for a classification system with a single feature that is independently and normally distributed for each class. Therefore, it may be of interest to consider a generalized approach for inference on BC when the feature used for classification is not normal (ex. gamma, chi square, mixtures, etc.). Finally, this work has assumed fixed prevalences on each class. However, it is possible that the prevalence of a class is not known explicitly. Future work may consider inference on BC when the prevalence of each class follows a known distribution to consider a possible range of prevalence values. Under this framework, Bayesian methods may be employed to determine properties of Bayes Cost as well as corresponding credible sets for Bayes Cost and the optimal thresholds.

Appendix A: Mathematical Derivations and Support

A.1 Asymptotic Distribution of Sample Mean and Variance

In order to show $(\overline{X}_n, S_n^2) \xrightarrow{d} mvn$, some necessary theorems and definition are presented first.

Definition 8 (Converges in Probability)**.**

A sequence of random variables, $X_1, X_2, \ldots$, converges in probability to a random variable X if for every $\epsilon > 0$, $\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$ or, equivalently, $\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1$ [12, p. 232]

Theorem 9 (Central Limit Theorem (CLT))**.**

Let $X_1, X_2, \ldots$ be a sequence of iid random variables with $EX_i = \mu$ and $0 < VarX_i = \sigma^2 < \infty$. Define $\overline{X}_n = (1/n)\sum_{i=1}^n X_i$. Let $G_n(x)$ denote the cdf of $\sqrt{n}(\overline{X}_n - \mu)/\sigma$. Then, for any x, $-\infty < x < \infty$,

$$\lim_{n\to\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

that is, $\sqrt{n}(\overline{X}_n - \mu)/\sigma$ has a limiting standard normal distribution. [12, p. 238]

Theorem 10 (Slutsky's Theorem)**.**

If $X_n \to X$ in distribution and $Y_n \to a$, a constant, in probability, then
a. $Y_n X_n \to aX$ in distribution
b. $X_n + Y_n \to X + a$ in distribution. [12, pg. 239-240]

Theorem 11.

Let $X_1, X_2, \ldots$, be iid $f(x \mid \theta)$, let $\widehat{\theta}$ denote the MLE of θ , and let $\tau(\theta)$ be a continuous function of θ . Under the regularity conditions [. . .] on $f(x \mid \theta)$ and , hence, $L(\theta \mid x)$,

$$\sqrt{n}[\tau(\widehat{\theta}) - \tau(\theta)] \to n[0, v(\theta)]$$

where $v(\theta)$ is the Cramér-Rao Lower Bound. That is, $\tau(\widehat{\theta})$ is a consisten and asymptotically efficient estimator of $\tau(\theta)$ [12, pg. 472]

Regularity conditions are presented in Section A.5, and are assumed for the normal distribution. From the CLT it is clear that $\overline{X}_n \xrightarrow{d} n[\mu, \sigma^2/n]$. To see that the sample variance (S_n^2) also has a limiting normal distribution first note from Theorem 11 that $\sqrt{n}[\widehat{\sigma}_n^2 - \sigma^2] \xrightarrow{d} n[0, v(\sigma^2)]$, where $\widehat{\sigma}^2$ is the Maximum Likelihood Estimator (MLE) of σ^2 and

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{n-1}{n} S_n^2 \tag{A.1}$$

Also, it is clear that

$$\lim_{n\to\infty} P\left(\left|\frac{\sqrt{n}}{n-1}\widehat{\sigma}_n^2 - 0\right| < \epsilon\right) = 1 \tag{A.2}$$

which implies from Definition 8 that $\frac{\sqrt{n}}{n-1}\widehat{\sigma}_n^2 \xrightarrow{p} 0$. Now consider,

$$\begin{aligned}\sqrt{n}\left(S_n^2-\sigma^2\right) &= \sqrt{n}\left(\frac{n}{n-1}\widehat{\sigma}_n^2-\sigma^2\right)\\ &= \sqrt{n}\left(\frac{n}{n-1}\widehat{\sigma}_n^2-\frac{1}{n-1}\widehat{\sigma}_n^2-\sigma^2\right)+\frac{\sqrt{n}}{n-1}\widehat{\sigma}_n^2\\ &= \sqrt{n}\left(\widehat{\sigma}_n^2-\sigma^2\right)+\frac{\sqrt{n}}{n-1}\widehat{\sigma}_n^2 \qquad \text{(A.3)}\end{aligned}$$

Let $X_n = \sqrt{n}\left(\widehat{\sigma}_n^2-\sigma^2\right)$ and $Y_n = \frac{\sqrt{n}}{n-1}\widehat{\sigma}_n^2$, then from Slutsky's Thm, $\sqrt{n}\left(S_n^2-\sigma^2\right) = X_n + Y_n \xrightarrow{d}$ $n[0, v(\sigma^2)] + 0 = n[0, v(\sigma^2)]$.

Finally, since $\overline{X}_n$ and S_n^2 are indpendent ([12, p. 218]), their asymptotic joint distribution is simply the product of their asymptotic normal marginals, which is the bivariate normal pdf with correlation, ρ , of zero. Therefore, $(\overline{X}_n\ ,\ S_n^2) \xrightarrow{d} mvn[(\mu,\sigma^2),(\sigma^2/n, v(\sigma^2))]$.

A.2 Derivation of partial derivatives of three-class Bayes Cost with respect to all distributional parameters.

$$\begin{aligned}\left(\frac{\partial BC}{\partial \mu_1}\right) &= \frac{\partial}{\partial \mu_1}\begin{bmatrix} c_{2|1}p_1\times\left(\Phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)-\Phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\right)+c_{3|1}p_1\times\left(\Phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\right)+\\ c_{1|2}p_2\times\left(\Phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\right)+c_{3|2}p_2\times\left(\Phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\right)+\\ c_{1|3}p_3\times\left(\Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right)+c_{2|3}p_3\times\left(\Phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)-\Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right)\end{bmatrix}\\ &= \begin{bmatrix} c_{2|1}p_1\times\frac{\partial}{\partial\mu_1}\left[\Phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\right]-c_{2|1}p_1\times\frac{\partial}{\partial\mu_1}\left[\Phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\right]+c_{3|1}p_1\times\frac{\partial}{\partial\mu_1}\left[\Phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\right]+\\ c_{1|2}p_2\times\frac{\partial}{\partial\mu_1}\left[\Phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\right]+c_{3|2}p_2\times\frac{\partial}{\partial\mu_1}\left[\Phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\right]+\\ c_{1|3}p_3\times\frac{\partial}{\partial\mu_1}\left[\Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right]+c_{2|3}p_3\times\frac{\partial}{\partial\mu_1}\left[\Phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\right]-c_{2|3}p_3\times\frac{\partial}{\partial\mu_1}\left[\Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right]\end{bmatrix}\\ &= \begin{bmatrix} c_{2|1}p_1\times\phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\frac{\partial}{\partial\mu_1}\left[\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\right]-c_{2|1}p_1\times\phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\frac{\partial}{\partial\mu_1}\left[\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\right]+\\ c_{3|1}p_1\times\phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\frac{\partial}{\partial\mu_1}\left[\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\right]+c_{1|2}p_2\times\phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\frac{\partial}{\partial\mu_1}\left[\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\right]+\\ c_{3|2}p_2\times\phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\frac{\partial}{\partial\mu_1}\left[\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\right]+c_{1|3}p_3\times\phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\frac{\partial}{\partial\mu_1}\left[\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right]+\\ c_{2|3}p_3\times\phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\frac{\partial}{\partial\mu_1}\left[\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\right]-c_{2|3}p_3\times\phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\frac{\partial}{\partial\mu_1}\left[\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right]\end{bmatrix}\\ &= \begin{bmatrix} c_{2|1}p_1\times\phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_2}{\partial\mu_1}\sigma_1^{-1}-\sigma_1^{-1}\right]-c_{2|1}p_1\times\phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_1}{\partial\mu_1}\sigma_1^{-1}-\sigma_1^{-1}\right]+\\ c_{3|1}p_1\times\phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\left[\sigma_1^{-1}-\frac{\partial\theta_2}{\partial\mu_1}\sigma_1^{-1}\right]+c_{1|2}p_2\times\phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial\theta_1}{\partial\mu_1}\sigma_2^{-1}\right]+\\ c_{3|2}p_2\times\phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[-\frac{\partial\theta_2}{\partial\mu_1}\sigma_2^{-1}\right]+c_{1|3}p_3\times\phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_1}{\partial\mu_1}\sigma_3^{-1}\right]+\\ c_{2|3}p_3\times\phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_2}{\partial\mu_1}\sigma_3^{-1}\right]-c_{2|3}p_3\times\phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_1}{\partial\mu_1}\sigma_3^{-1}\right]\end{bmatrix}\end{aligned}$$

Pulling out the standard deviations and using $\phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right) = \phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)$ results in

$$\left(\frac{\partial BC}{\partial \mu_1}\right) = \begin{bmatrix} \sigma_1^{-1}\left[\begin{array}{c} c_{2|1}p_1\phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_2}{\partial\mu_1}-1\right] - c_{2|1}p_1\phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_1}{\partial\mu_1}-1\right]+ \\ c_{3|1}p_1\phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\left[1-\frac{\partial\theta_2}{\partial\mu_1}\right] \end{array}\right]+ \\ \sigma_2^{-1}\left[c_{1|2}p_2\phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial\theta_1}{\partial\mu_1}\right] + c_{3|2}p_2\phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[-\frac{\partial\theta_2}{\partial\mu_1}\right]\right]+ \\ \sigma_3^{-1}\left[c_{1|3}p_3\phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_1}{\partial\mu_1}\right] + c_{2|3}p_3\phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_2}{\partial\mu_1}\right] - c_{2|3}p_3\phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_1}{\partial\mu_1}\right]\right] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^{-1}\left[\left[\frac{\partial\theta_2}{\partial\mu_1}-1\right]\left(c_{2|1}p_1\phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right) - c_{3|1}p_1\phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\right) - c_{2|1}p_1\phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_1}{\partial\mu_1}-1\right]\right]+ \\ \sigma_2^{-1}\left[c_{1|2}p_2\phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial\theta_1}{\partial\mu_1}\right] + c_{3|2}p_2\phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[-\frac{\partial\theta_2}{\partial\mu_1}\right]\right]+ \\ \sigma_3^{-1}\left[\left[\frac{\partial\theta_1}{\partial\mu_1}\right]\left(c_{1|3}p_3\phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right) - c_{2|3}p_3\phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right) + c_{2|3}p_3\phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_2}{\partial\mu_1}\right]\right] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^{-1}\left[\left[\frac{\partial\theta_2}{\partial\mu_1}-1\right]\phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)(c_{2|1}p_1 - c_{3|1}p_1) - c_{2|1}p_1\phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_1}{\partial\mu_1}-1\right]\right]+ \\ \sigma_2^{-1}\left[c_{1|2}p_2\phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial\theta_1}{\partial\mu_1}\right] + c_{3|2}p_2\phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[-\frac{\partial\theta_2}{\partial\mu_1}\right]\right]+ \\ \sigma_3^{-1}\left[\left[\frac{\partial\theta_1}{\partial\mu_1}\right]\phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)(c_{1|3}p_3 - c_{2|3}p_3) + c_{2|3}p_3\phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_2}{\partial\mu_1}\right]\right] \end{bmatrix}$$

$$\left(\frac{\partial BC}{\partial \mu_2}\right) = \frac{\partial}{\partial\mu_2}\begin{bmatrix} c_{2|1}p_1 \times \left(\Phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right) - \Phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\right) + c_{3|1}p_1 \times \left(\Phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\right)+ \\ c_{1|2}p_2 \times \left(\Phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\right) + c_{3|2}p_2 \times \left(\Phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\right)+ \\ c_{1|3}p_3 \times \left(\Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right) + c_{2|3}p_3 \times \left(\Phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right) - \Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right) \end{bmatrix}$$

$$= \begin{bmatrix} c_{2|1}p_1 \times \frac{\partial}{\partial\mu_2}\left[\Phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\right] - c_{2|1}p_1 \times \frac{\partial}{\partial\mu_2}\left[\Phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\right] + c_{3|1}p_1 \times \frac{\partial}{\partial\mu_2}\left[\Phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\right]+ \\ c_{1|2}p_2 \times \frac{\partial}{\partial\mu_2}\left[\Phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\right] + c_{3|2}p_2 \times \frac{\partial}{\partial\mu_2}\left[\Phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\right]+ \\ c_{1|3}p_3 \times \frac{\partial}{\partial\mu_2}\left[\Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right] + c_{2|3}p_3 \times \frac{\partial}{\partial\mu_2}\left[\Phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\right] - c_{2|3}p_3 \times \frac{\partial}{\partial\mu_2}\left[\Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right] \end{bmatrix}$$

$$= \begin{bmatrix} c_{2|1}p_1 \times \phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\frac{\partial}{\partial\mu_2}\left[\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\right] - c_{2|1}p_1 \times \phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\frac{\partial}{\partial\mu_2}\left[\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\right]+ \\ c_{3|1}p_1 \times \phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\frac{\partial}{\partial\mu_2}\left[\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\right] + c_{1|2}p_2 \times \phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\frac{\partial}{\partial\mu_2}\left[\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\right]+ \\ c_{3|2}p_2 \times \phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\frac{\partial}{\partial\mu_2}\left[\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\right] + c_{1|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\frac{\partial}{\partial\mu_2}\left[\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right]+ \\ c_{2|3}p_3 \times \phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\frac{\partial}{\partial\mu_2}\left[\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\right] - c_{2|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\frac{\partial}{\partial\mu_2}\left[\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right] \end{bmatrix}$$

$$= \begin{bmatrix} c_{2|1}p_1 \times \phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_2}{\partial\mu_2}\sigma_1^{-1}\right] - c_{2|1}p_1 \times \phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_1}{\partial\mu_2}\sigma_1^{-1}\right]+ \\ c_{3|1}p_1 \times \phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\left[-\frac{\partial\theta_2}{\partial\mu_2}\sigma_1^{-1}\right] + c_{1|2}p_2 \times \phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial\theta_1}{\partial\mu_2}\sigma_2^{-1} - \sigma_2^{-1}\right]+ \\ c_{3|2}p_2 \times \phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[\sigma_2^{-1} - \frac{\partial\theta_2}{\partial\mu_2}\sigma_2^{-1}\right] + c_{1|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_1}{\partial\mu_2}\sigma_3^{-1}\right]+ \\ c_{2|3}p_3 \times \phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_2}{\partial\mu_2}\sigma_3^{-1}\right] - c_{2|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_1}{\partial\mu_2}\sigma_3^{-1}\right] \end{bmatrix}$$

Pulling out the standard deviations and using $\phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right) = \phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)$ results in

$$
\begin{aligned}
\left(\frac{\partial BC}{\partial \mu_2}\right) &= \begin{bmatrix}
\sigma_1^{-1}\left[c_{2|1}p_1\phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_2}{\partial\mu_2}\right] - c_{2|1}p_1\phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_1}{\partial\mu_2}\right] + c_{3|1}p_1\phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\left[-\frac{\partial\theta_2}{\partial\mu_2}\right]\right] + \\
\sigma_2^{-1}\left[c_{1|2}p_2\phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial\theta_1}{\partial\mu_2} - 1\right] + c_{3|2}p_2\phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[1 - \frac{\partial\theta_2}{\partial\mu_2}\right]\right] + \\
\sigma_3^{-1}\left[c_{1|3}p_3\phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_1}{\partial\mu_2}\right] + c_{2|3}p_3\phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_2}{\partial\mu_2}\right] - c_{2|3}p_3\phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_1}{\partial\mu_2}\right]\right]
\end{bmatrix} \\
&= \begin{bmatrix}
\sigma_1^{-1}\left[\left[\frac{\partial\theta_2}{\partial\mu_2}\right]\left(c_{2|1}p_1\phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right) - c_{3|1}p_1\phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\right) - c_{2|1}p_1\phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_1}{\partial\mu_2}\right]\right] + \\
\sigma_2^{-1}\left[c_{1|2}p_2\phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial\theta_1}{\partial\mu_2} - 1\right] + c_{3|2}p_2\phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[1 - \frac{\partial\theta_2}{\partial\mu_2}\right]\right] + \\
\sigma_3^{-1}\left[\left[\frac{\partial\theta_1}{\partial\mu_2}\right]\left(c_{1|3}p_3\phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right) - c_{2|3}p_3\phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right) + c_{2|3}p_3\phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_2}{\partial\mu_2}\right]\right]
\end{bmatrix} \\
&= \begin{bmatrix}
\sigma_1^{-1}\left[\left[\frac{\partial\theta_2}{\partial\mu_2}\right]\phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)(c_{2|1}p_1 - c_{3|1}p_1) - c_{2|1}p_1\phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_1}{\partial\mu_2}\right]\right] + \\
\sigma_2^{-1}\left[c_{1|2}p_2\phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial\theta_1}{\partial\mu_2} - 1\right] + c_{3|2}p_2\phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[1 - \frac{\partial\theta_2}{\partial\mu_2}\right]\right] + \\
\sigma_3^{-1}\left[\left[\frac{\partial\theta_1}{\partial\mu_2}\right]\phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)(c_{1|3}p_3 - c_{2|3}p_3) + c_{2|3}p_3\phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_2}{\partial\mu_2}\right]\right]
\end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\left(\frac{\partial BC}{\partial \mu_3}\right) &= \frac{\partial}{\partial \mu_3}\begin{bmatrix} c_{2|1}p_1 \times \left(\Phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right) - \Phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\right) + c_{3|1}p_1 \times \left(\Phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\right) + \\ c_{1|2}p_2 \times \left(\Phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\right) + c_{3|2}p_2 \times \left(\Phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\right) + \\ c_{1|3}p_3 \times \left(\Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right) + c_{2|3}p_3 \times \left(\Phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right) - \Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right) \end{bmatrix} \\
&= \begin{bmatrix} c_{2|1}p_1 \times \frac{\partial}{\partial \mu_3}\left[\Phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\right] - c_{2|1}p_1 \times \frac{\partial}{\partial \mu_3}\left[\Phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\right] + c_{3|1}p_1 \times \frac{\partial}{\partial \mu_3}\left[\Phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\right] + \\ c_{1|2}p_2 \times \frac{\partial}{\partial \mu_3}\left[\Phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\right] + c_{3|2}p_2 \times \frac{\partial}{\partial \mu_3}\left[\Phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\right] + \\ c_{1|3}p_3 \times \frac{\partial}{\partial \mu_3}\left[\Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right] + c_{2|3}p_3 \times \frac{\partial}{\partial \mu_3}\left[\Phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\right] - c_{2|3}p_3 \times \frac{\partial}{\partial \mu_3}\left[\Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right] \end{bmatrix} \\
&= \begin{bmatrix} c_{2|1}p_1 \times \phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\frac{\partial}{\partial \mu_3}\left[\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\right] - c_{2|1}p_1 \times \phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\frac{\partial}{\partial \mu_3}\left[\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\right] + \\ c_{3|1}p_1 \times \phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\frac{\partial}{\partial \mu_3}\left[\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\right] + c_{1|2}p_2 \times \phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\frac{\partial}{\partial \mu_3}\left[\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\right] + \\ c_{3|2}p_2 \times \phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\frac{\partial}{\partial \mu_3}\left[\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\right] + c_{1|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\frac{\partial}{\partial \mu_3}\left[\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right] + \\ c_{2|3}p_3 \times \phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\frac{\partial}{\partial \mu_3}\left[\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\right] - c_{2|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\frac{\partial}{\partial \mu_3}\left[\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right] \end{bmatrix} \\
&= \begin{bmatrix} c_{2|1}p_1 \times \phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_2}{\partial \mu_3}\sigma_1^{-1}\right] - \\ c_{2|1}p_1 \times \phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_1}{\partial \mu_3}\sigma_1^{-1}\right] + \\ c_{3|1}p_1 \times \phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\left[-\frac{\partial \theta_2}{\partial \mu_3}\sigma_1^{-1}\right] + \\ c_{1|2}p_2 \times \phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial \theta_1}{\partial \mu_3}\sigma_2^{-1}\right] + \\ c_{3|2}p_2 \times \phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[-\frac{\partial \theta_2}{\partial \mu_3}\sigma_2^{-1}\right] + \\ c_{1|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_1}{\partial \mu_3}\sigma_3^{-1} - \sigma_3^{-1}\right] + \\ c_{2|3}p_3 \times \phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_2}{\partial \mu_3}\sigma_3^{-1} - \sigma_3^{-1}\right] - \\ c_{2|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_1}{\partial \mu_3}\sigma_3^{-1} - \sigma_3^{-1}\right] \end{bmatrix} \\
&= \begin{bmatrix} \sigma_1^{-1}\left[c_{2|1}p_1\phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_2}{\partial \mu_3}\right] - c_{2|1}p_1\phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_1}{\partial \mu_3}\right] + c_{3|1}p_1\phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\left[-\frac{\partial \theta_2}{\partial \mu_3}\right]\right] + \\ \sigma_2^{-1}\left[c_{1|2}p_2\phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial \theta_1}{\partial \mu_3}\right] + c_{3|2}p_2\phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[-\frac{\partial \theta_2}{\partial \mu_3}\right]\right] + \\ \sigma_3^{-1}\left[c_{1|3}p_3\phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_1}{\partial \mu_3} - 1\right] + c_{2|3}p_3\phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_2}{\partial \mu_3} - 1\right] - c_{2|3}p_3\phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_1}{\partial \mu_3} - 1\right]\right] \end{bmatrix} \\
&= \begin{bmatrix} \sigma_1^{-1}\left[\left[\frac{\partial \theta_2}{\partial \mu_3}\right]\left(c_{2|1}p_1\phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right) - c_{3|1}p_1\phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\right) - c_{2|1}p_1\phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_1}{\partial \mu_3}\right]\right] + \\ \sigma_2^{-1}\left[c_{1|2}p_2\phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial \theta_1}{\partial \mu_3}\right] + c_{3|2}p_2\phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[-\frac{\partial \theta_2}{\partial \mu_3}\right]\right] + \\ \sigma_3^{-1}\left[\left[\frac{\partial \theta_1}{\partial \mu_3} - 1\right]\left(c_{1|3}p_3\phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right) - c_{2|3}p_3\phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right) + c_{2|3}p_3\phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_2}{\partial \mu_3} - 1\right]\right] \end{bmatrix} \\
&= \begin{bmatrix} \sigma_1^{-1}\left[\left[\frac{\partial \theta_2}{\partial \mu_3}\right]\phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)(c_{2|1}p_1 - c_{3|1}p_1) - c_{2|1}p_1\phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_1}{\partial \mu_3}\right]\right] + \\ \sigma_2^{-1}\left[c_{1|2}p_2\phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial \theta_1}{\partial \mu_3}\right] + c_{3|2}p_2\phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[-\frac{\partial \theta_2}{\partial \mu_3}\right]\right] + \\ \sigma_3^{-1}\left[\left[\frac{\partial \theta_1}{\partial \mu_3} - 1\right]\phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)(c_{1|3}p_3 - c_{2|3}p_3) + c_{2|3}p_3\phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_2}{\partial \mu_3} - 1\right]\right] \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\left(\frac{\partial BC}{\partial \sigma_1}\right) &= \frac{\partial}{\partial \sigma_1}\begin{bmatrix} c_{2|1}p_1 \times \left(\Phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right) - \Phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\right) + c_{3|1}p_1 \times \left(\Phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\right) + \\ c_{1|2}p_2 \times \left(\Phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\right) + c_{3|2}p_2 \times \left(\Phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\right) + \\ c_{1|3}p_3 \times \left(\Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right) + c_{2|3}p_3 \times \left(\Phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right) - \Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right) \end{bmatrix} \\
&= \begin{bmatrix} c_{2|1}p_1 \times \frac{\partial}{\partial \sigma_1}\left[\Phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\right] - c_{2|1}p_1 \times \frac{\partial}{\partial \sigma_1}\left[\Phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\right] + c_{3|1}p_1 \times \frac{\partial}{\partial \sigma_1}\left[\Phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\right] + \\ c_{1|2}p_2 \times \frac{\partial}{\partial \sigma_1}\left[\Phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\right] + c_{3|2}p_2 \times \frac{\partial}{\partial \sigma_1}\left[\Phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\right] + \\ c_{1|3}p_3 \times \frac{\partial}{\partial \sigma_1}\left[\Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right] + c_{2|3}p_3 \times \frac{\partial}{\partial \sigma_1}\left[\Phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\right] - c_{2|3}p_3 \times \frac{\partial}{\partial \sigma_1}\left[\Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right] \end{bmatrix} \\
&= \begin{bmatrix} c_{2|1}p_1 \times \phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\frac{\partial}{\partial \sigma_1}\left[\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\right] - c_{2|1}p_1 \times \phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\frac{\partial}{\partial \sigma_1}\left[\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\right] + \\ c_{3|1}p_1 \times \phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\frac{\partial}{\partial \sigma_1}\left[\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\right] + c_{1|2}p_2 \times \phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\frac{\partial}{\partial \sigma_1}\left[\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\right] + \\ c_{3|2}p_2 \times \phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\frac{\partial}{\partial \sigma_1}\left[\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\right] + c_{1|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\frac{\partial}{\partial \sigma_1}\left[\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right] + \\ c_{2|3}p_3 \times \phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\frac{\partial}{\partial \sigma_1}\left[\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\right] - c_{2|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\frac{\partial}{\partial \sigma_1}\left[\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right] \end{bmatrix}
\end{aligned}
$$

$$
\left(\frac{\partial BC}{\partial \sigma_1}\right) = \begin{bmatrix}
c_{2|1}p_1 \times \phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\left[\frac{1}{\sigma_1}\frac{\partial}{\partial\sigma_1}(\theta_2-\mu_1) + (\theta_2-\mu_1)\frac{\partial}{\partial\sigma_1}(\frac{1}{\sigma_1})\right] - \\
c_{2|1}p_1 \times \phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{1}{\sigma_1}\frac{\partial}{\partial\sigma_1}(\theta_1-\mu_1) + (\theta_1-\mu_1)\frac{\partial}{\partial\sigma_1}(\frac{1}{\sigma_1})\right] + \\
c_{3|1}p_1 \times \phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\left[\frac{1}{\sigma_1}\frac{\partial}{\partial\sigma_1}(\mu_1-\theta_2) + (\mu_1-\theta_2)\frac{\partial}{\partial\sigma_1}\frac{1}{\sigma_1}\right] + \\
c_{1|2}p_2 \times \phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial\theta_1}{\partial\sigma_1}\sigma_2^{-1}\right] + \\
c_{3|2}p_2 \times \phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[-\frac{\partial\theta_2}{\partial\sigma_1}\sigma_2^{-1}\right] + \\
c_{1|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_1}{\partial\sigma_1}\sigma_3^{-1}\right] + \\
c_{2|3}p_3 \times \phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_2}{\partial\sigma_1}\sigma_3^{-1}\right] - \\
c_{2|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_1}{\partial\sigma_1}\sigma_3^{-1}\right]
\end{bmatrix}
$$

$$
= \begin{bmatrix}
c_{2|1}p_1 \times \phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\left[\sigma_1^{-1}\frac{\partial\theta_2}{\partial\sigma_1} - (\theta_2-\mu_1)\sigma_1^{-2}\right] - \\
c_{2|1}p_1 \times \phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{1}{\sigma_1}\frac{\partial\theta_1}{\partial\sigma_1} - (\theta_1-\mu_1)\sigma_1^{-2}\right] + \\
c_{3|1}p_1 \times \phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\left[\frac{1}{\sigma_1}\frac{-\partial\theta_2}{\partial\sigma_1} - (\mu_1-\theta_2)\sigma_1^{-2}\right] + \\
c_{1|2}p_2 \times \phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial\theta_1}{\partial\sigma_1}\sigma_2^{-1}\right] + \\
c_{3|2}p_2 \times \phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[-\frac{\partial\theta_2}{\partial\sigma_1}\sigma_2^{-1}\right] + \\
c_{1|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_1}{\partial\sigma_1}\sigma_3^{-1}\right] + \\
c_{2|3}p_3 \times \phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_2}{\partial\sigma_1}\sigma_3^{-1}\right] - \\
c_{2|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_1}{\partial\sigma_1}\sigma_3^{-1}\right]
\end{bmatrix}
$$

$$
= \begin{bmatrix}
\sigma_1^{-1}c_{2|1}p_1 \times \phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_2}{\partial\sigma_1} - (\frac{\theta_2-\mu_1}{\sigma_1})\right] - \\
\sigma_1^{-1}c_{2|1}p_1 \times \phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_1}{\partial\sigma_1} - (\frac{\theta_1-\mu_1}{\sigma_1})\right] + \\
\sigma_1^{-1}c_{3|1}p_1 \times \phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\left[\frac{-\partial\theta_2}{\partial\sigma_1} - (\frac{\mu_1-\theta_2}{\sigma_1})\right] + \\
\sigma_2^{-1}c_{1|2}p_2 \times \phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial\theta_1}{\partial\sigma_1}\right] + \\
\sigma_2^{-1}c_{3|2}p_2 \times \phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[-\frac{\partial\theta_2}{\partial\sigma_1}\right] + \\
\sigma_3^{-1}c_{1|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_1}{\partial\sigma_1}\right] + \\
\sigma_3^{-1}c_{2|3}p_3 \times \phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_2}{\partial\sigma_1}\right] - \\
\sigma_3^{-1}c_{2|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_1}{\partial\sigma_1}\right]
\end{bmatrix}
$$

and continuing to simplify:

$$
\begin{aligned}
\left(\frac{\partial BC}{\partial \sigma_1}\right) &= \begin{bmatrix}
\sigma_1^{-1} c_{2|1} p_1 \times \phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_2}{\partial \sigma_1} - (\frac{\theta_2-\mu_1}{\sigma_1})\right] - \\
\sigma_1^{-1} c_{2|1} p_1 \times \phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_1}{\partial \sigma_1} - (\frac{\theta_1-\mu_1}{\sigma_1})\right] + \\
\sigma_1^{-1} c_{3|1} p_1 \times \phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\left[\frac{-\partial \theta_2}{\partial \sigma_1} - (\frac{\mu_1-\theta_2}{\sigma_1})\right] + \\
\sigma_2^{-1} c_{1|2} p_2 \times \phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial \theta_1}{\partial \sigma_1}\right] + \\
\sigma_2^{-1} c_{3|2} p_2 \times \phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[-\frac{\partial \theta_2}{\partial \sigma_1}\right] + \\
\sigma_3^{-1} c_{1|3} p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_1}{\partial \sigma_1}\right] + \\
\sigma_3^{-1} c_{2|3} p_3 \times \phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_2}{\partial \sigma_1}\right] - \\
\sigma_3^{-1} c_{2|3} p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_1}{\partial \sigma_1}\right]
\end{bmatrix} \\
&= \begin{bmatrix}
\sigma_1^{-1}\left[\begin{array}{c} c_{2|1} p_1 \phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_2}{\partial \sigma_1} - (\frac{\theta_2-\mu_1}{\sigma_1})\right] - c_{2|1} p_1 \phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_1}{\partial \sigma_1} - (\frac{\theta_1-\mu_1}{\sigma_1})\right] + \\ c_{3|1} p_1 \phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\left[\frac{-\partial \theta_2}{\partial \sigma_1} - (\frac{\mu_1-\theta_2}{\sigma_1})\right] \end{array}\right] + \\
\sigma_2^{-1}\left[c_{1|2} p_2 \phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial \theta_1}{\partial \sigma_1}\right] + c_{3|2} p_2 \phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[-\frac{\partial \theta_2}{\partial \sigma_1}\right]\right] + \\
\sigma_3^{-1}\left[c_{1|3} p_3 \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_1}{\partial \sigma_1}\right] + c_{2|3} p_3 \phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_2}{\partial \sigma_1}\right] - c_{2|3} p_3 \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_1}{\partial \sigma_1}\right]\right]
\end{bmatrix} \\
&= \begin{bmatrix}
\sigma_1^{-1}\left[\phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)(c_{2|1} p_1 - c_{3|1} p_1)\left[\frac{\partial \theta_2}{\partial \sigma_1} - (\frac{\theta_2-\mu_1}{\sigma_1})\right] - c_{2|1} p_1 \phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_1}{\partial \sigma_1} - (\frac{\theta_1-\mu_1}{\sigma_1})\right]\right] + \\
\sigma_2^{-1}\left[c_{1|2} p_2 \phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial \theta_1}{\partial \sigma_1}\right] - c_{3|2} p_2 \phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[\frac{\partial \theta_2}{\partial \sigma_1}\right]\right] + \\
\sigma_3^{-1}\left[\phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_1}{\partial \sigma_1}\right](c_{1|3} p_3 - c_{2|3} p_3) + c_{2|3} p_3 \phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_2}{\partial \sigma_1}\right]\right]
\end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\left(\frac{\partial BC}{\partial \sigma 2}\right) &= \frac{\partial}{\partial \sigma_2}\begin{bmatrix} c_{2|1}p_1 \times \left(\Phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right) - \Phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\right) + c_{3|1}p_1 \times \left(\Phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\right) + \\ c_{1|2}p_2 \times \left(\Phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\right) + c_{3|2}p_2 \times \left(\Phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\right) + \\ c_{1|3}p_3 \times \left(\Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right) + c_{2|3}p_3 \times \left(\Phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right) - \Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right) \end{bmatrix} \\
&= \begin{bmatrix} c_{2|1}p_1 \times \frac{\partial}{\partial\sigma_2}\left[\Phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\right] - c_{2|1}p_1 \times \frac{\partial}{\partial\sigma_2}\left[\Phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\right] + c_{3|1}p_1 \times \frac{\partial}{\partial\sigma_2}\left[\Phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\right] + \\ c_{1|2}p_2 \times \frac{\partial}{\partial\sigma_2}\left[\Phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\right] + c_{3|2}p_2 \times \frac{\partial}{\partial\sigma_2}\left[\Phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\right] + \\ c_{1|3}p_3 \times \frac{\partial}{\partial\sigma_2}\left[\Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right] + c_{2|3}p_3 \times \frac{\partial}{\partial\sigma_2}\left[\Phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\right] - c_{2|3}p_3 \times \frac{\partial}{\partial\sigma_2}\left[\Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right] \end{bmatrix} \\
&= \begin{bmatrix} c_{2|1}p_1 \times \phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\frac{\partial}{\partial\sigma_2}\left[\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\right] - c_{2|1}p_1 \times \phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\frac{\partial}{\partial\sigma_2}\left[\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\right] + \\ c_{3|1}p_1 \times \phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\frac{\partial}{\partial\sigma_2}\left[\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\right] + c_{1|2}p_2 \times \phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\frac{\partial}{\partial\sigma_2}\left[\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\right] + \\ c_{3|2}p_2 \times \phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\frac{\partial}{\partial\sigma_2}\left[\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\right] + c_{1|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\frac{\partial}{\partial\sigma_2}\left[\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right] + \\ c_{2|3}p_3 \times \phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\frac{\partial}{\partial\sigma_2}\left[\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\right] - c_{2|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\frac{\partial}{\partial\sigma_2}\left[\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right] \end{bmatrix} \\
&= \begin{bmatrix} c_{2|1}p_1 \times \phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_2}{\partial\sigma_2}\sigma_1^{-1}\right] - \\ c_{2|1}p_1 \times \phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_1}{\partial\sigma_2}\sigma_1^{-1}\right] + \\ c_{3|1}p_1 \times \phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\left[-\frac{\partial\theta_2}{\partial\sigma_2}\sigma_1^{-1}\right] + \\ c_{1|2}p_2 \times \phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{1}{\sigma_2}\frac{\partial}{\partial\sigma_2}(\theta_1-\mu_2) + (\theta_1-\mu_2)\frac{\partial}{\partial\sigma_2}\left(\frac{1}{\sigma_2}\right)\right] + \\ c_{3|2}p_2 \times \phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[\frac{1}{\sigma_2}\frac{\partial}{\partial\sigma_2}(\mu_2-\theta_2) + (\mu_2-\theta_2)\frac{\partial}{\partial\sigma_2}\left(\frac{1}{\sigma_2}\right)\right] + \\ c_{1|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_1}{\partial\sigma_2}\sigma_3^{-1}\right] + \\ c_{2|3}p_3 \times \phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_2}{\partial\sigma_2}\sigma_3^{-1}\right] - \\ c_{2|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_1}{\partial\sigma_2}\sigma_3^{-1}\right] \end{bmatrix} \\
&= \begin{bmatrix} c_{2|1}p_1 \times \phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_2}{\partial\sigma_2}\sigma_1^{-1}\right] - \\ c_{2|1}p_1 \times \phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_1}{\partial\sigma_2}\sigma_1^{-1}\right] + \\ c_{3|1}p_1 \times \phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\left[-\frac{\partial\theta_2}{\partial\sigma_2}\sigma_1^{-1}\right] + \\ c_{1|2}p_2 \times \phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\sigma_2^{-1}\frac{\partial\theta_1}{\partial\sigma_2} - (\theta_1-\mu_2)\sigma_2^{-2}\right] + \\ c_{3|2}p_2 \times \phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[\sigma_2^{-1}\frac{-\partial\theta_2}{\partial\sigma_2} - (\mu_2-\theta_2)\sigma_2^{-2}\right] + \\ c_{1|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_1}{\partial\sigma_2}\sigma_3^{-1}\right] + \\ c_{2|3}p_3 \times \phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_2}{\partial\sigma_2}\sigma_3^{-1}\right] - \\ c_{2|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_1}{\partial\sigma_2}\sigma_3^{-1}\right] \end{bmatrix}
\end{aligned}
$$

and continuing to simplify:

$$
\begin{aligned}
\left(\frac{\partial BC}{\partial \sigma_2}\right) &= \begin{bmatrix}
\sigma_1^{-1} c_{2|1} p_1 \times \phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_2}{\partial \sigma_2}\right] - \\
\sigma_1^{-1} c_{2|1} p_1 \times \phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_1}{\partial \sigma_2}\right] + \\
\sigma_1^{-1} c_{3|1} p_1 \times \phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\left[-\frac{\partial \theta_2}{\partial \sigma_2}\right] + \\
\sigma_2^{-1} c_{1|2} p_2 \times \phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial \theta_1}{\partial \sigma_2} - \left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\right] + \\
\sigma_2^{-1} c_{3|2} p_2 \times \phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[\frac{-\partial \theta_2}{\partial \sigma_2} - \left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\right] + \\
\sigma_3^{-1} c_{1|3} p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_1}{\partial \sigma_2}\right] + \\
\sigma_3^{-1} c_{2|3} p_3 \times \phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_2}{\partial \sigma_2}\right] - \\
\sigma_3^{-1} c_{2|3} p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_1}{\partial \sigma_2}\right]
\end{bmatrix} \\
&= \begin{bmatrix}
\sigma_1^{-1} c_{2|1} p_1 \times \phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_2}{\partial \sigma_2}\right] - \\
\sigma_1^{-1} c_{2|1} p_1 \times \phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_1}{\partial \sigma_2}\right] + \\
\sigma_1^{-1} c_{3|1} p_1 \times \phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\left[-\frac{\partial \theta_2}{\partial \sigma_2}\right] + \\
\sigma_2^{-1} c_{1|2} p_2 \times \phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial \theta_1}{\partial \sigma_2} - \left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\right] + \\
\sigma_2^{-1} c_{3|2} p_2 \times \phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[\frac{-\partial \theta_2}{\partial \sigma_2} - \left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\right] + \\
\sigma_3^{-1} c_{1|3} p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_1}{\partial \sigma_2}\right] + \\
\sigma_3^{-1} c_{2|3} p_3 \times \phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_2}{\partial \sigma_2}\right] - \\
\sigma_3^{-1} c_{2|3} p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_1}{\partial \sigma_2}\right]
\end{bmatrix} \\
&= \begin{bmatrix}
\sigma_1^{-1}\left[c_{2|1} p_1 \phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_2}{\partial \sigma_2}\right] - c_{2|1} p_1 \phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_1}{\partial \sigma_2}\right] + c_{3|1} p_1 \phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\left[-\frac{\partial \theta_2}{\partial \sigma_2}\right]\right] + \\
\sigma_2^{-1}\left[c_{1|2} p_2 \phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial \theta_1}{\partial \sigma_2} - \left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\right] + c_{3|2} p_2 \phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[\frac{-\partial \theta_2}{\partial \sigma_2} - \left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\right]\right] + \\
\sigma_3^{-1}\left[c_{1|3} p_3 \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_1}{\partial \sigma_2}\right] + c_{2|3} p_3 \phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_2}{\partial \sigma_2}\right] - c_{2|3} p_3 \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_1}{\partial \sigma_2}\right]\right]
\end{bmatrix} \\
&= \begin{bmatrix}
\sigma_1^{-1}\left[\phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_2}{\partial \sigma_2}\right](c_{2|1} p_1 - c_{3|1} p_1) - c_{2|1} p_1 \phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_1}{\partial \sigma_2}\right]\right] + \\
\sigma_2^{-1}\left[c_{1|2} p_2 \phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial \theta_1}{\partial \sigma_2} - \left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\right] + c_{3|2} p_2 \phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[\frac{-\partial \theta_2}{\partial \sigma_2} - \left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\right]\right] + \\
\sigma_3^{-1}\left[\phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_1}{\partial \sigma_2}\right](c_{1|3} p_3 - c_{2|3} p_3) + c_{2|3} p_3 \phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_2}{\partial \sigma_2}\right]\right]
\end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\left(\frac{\partial BC}{\partial \sigma_3}\right) &= \frac{\partial}{\partial \sigma_3}\begin{bmatrix} c_{2|1}p_1 \times \left(\Phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right) - \Phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\right) + c_{3|1}p_1 \times \left(\Phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\right) + \\ c_{1|2}p_2 \times \left(\Phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\right) + c_{3|2}p_2 \times \left(\Phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\right) + \\ c_{1|3}p_3 \times \left(\Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right) + c_{2|3}p_3 \times \left(\Phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right) - \Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right) \end{bmatrix} \\
&= \begin{bmatrix} c_{2|1}p_1 \times \frac{\partial}{\partial \sigma_3}\left[\Phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\right] - c_{2|1}p_1 \times \frac{\partial}{\partial \sigma_3}\left[\Phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\right] + c_{3|1}p_1 \times \frac{\partial}{\partial \sigma_3}\left[\Phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\right] + \\ c_{1|2}p_2 \times \frac{\partial}{\partial \sigma_3}\left[\Phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\right] + c_{3|2}p_2 \times \frac{\partial}{\partial \sigma_3}\left[\Phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\right] + \\ c_{1|3}p_3 \times \frac{\partial}{\partial \sigma_3}\left[\Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right] + c_{2|3}p_3 \times \frac{\partial}{\partial \sigma_3}\left[\Phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\right] - c_{2|3}p_3 \times \frac{\partial}{\partial \sigma_3}\left[\Phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right] \end{bmatrix} \\
&= \begin{bmatrix} c_{2|1}p_1 \times \phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\frac{\partial}{\partial \sigma_3}\left[\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\right] - c_{2|1}p_1 \times \phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\frac{\partial}{\partial \sigma_3}\left[\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\right] + \\ c_{3|1}p_1 \times \phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\frac{\partial}{\partial \sigma_3}\left[\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\right] + c_{1|2}p_2 \times \phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\frac{\partial}{\partial \sigma_3}\left[\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\right] + \\ c_{3|2}p_2 \times \phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\frac{\partial}{\partial \sigma_3}\left[\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\right] + c_{1|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\frac{\partial}{\partial \sigma_3}\left[\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right] + \\ c_{2|3}p_3 \times \phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\frac{\partial}{\partial \sigma_3}\left[\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\right] - c_{2|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\frac{\partial}{\partial \sigma_3}\left[\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right] \end{bmatrix} \\
&= \begin{bmatrix} c_{2|1}p_1 \times \phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_2}{\partial \sigma_3}\sigma_1^{-1}\right] - \\ c_{2|1}p_1 \times \phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_1}{\partial \sigma_3}\sigma_1^{-1}\right] + \\ c_{3|1}p_1 \times \phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\left[-\frac{\partial \theta_2}{\partial \sigma_3}\sigma_1^{-1}\right] + \\ c_{1|2}p_2 \times \phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial \theta_1}{\partial \sigma_3}\sigma_2^{-1}\right] + \\ c_{3|2}p_2 \times \phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[-\frac{\partial \theta_2}{\partial \sigma_3}\sigma_2^{-1}\right] + \\ c_{1|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{1}{\sigma_3}\frac{\partial}{\partial \sigma_3}(\theta_1-\mu_3) + (\theta_1-\mu_3)\frac{\partial}{\partial \sigma_3}\left(\frac{1}{\sigma_3}\right)\right] + \\ c_{2|3}p_3 \times \phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{1}{\sigma_3}\frac{\partial}{\partial \sigma_3}(\theta_2-\mu_3) + (\theta_2-\mu_3)\frac{\partial}{\partial \sigma_3}\left(\frac{1}{\sigma_3}\right)\right] - \\ c_{2|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{1}{\sigma_3}\frac{\partial}{\partial \sigma_3}(\theta_1-\mu_3) + (\theta_1-\mu_3)\frac{\partial}{\partial \sigma_3}\left(\frac{1}{\sigma_3}\right)\right] \end{bmatrix} \\
&= \begin{bmatrix} c_{2|1}p_1 \times \phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_2}{\partial \sigma_3}\sigma_1^{-1}\right] - \\ c_{2|1}p_1 \times \phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_1}{\partial \sigma_3}\sigma_1^{-1}\right] + \\ c_{3|1}p_1 \times \phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\left[-\frac{\partial \theta_2}{\partial \sigma_3}\sigma_1^{-1}\right] + \\ c_{1|2}p_2 \times \phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial \theta_1}{\partial \sigma_3}\sigma_2^{-1}\right] + \\ c_{3|2}p_2 \times \phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[-\frac{\partial \theta_2}{\partial \sigma_3}\sigma_2^{-1}\right] + \\ c_{1|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\sigma_3^{-1}\frac{\partial \theta_1}{\partial \sigma_3} - (\theta_1-\mu_3)\sigma_3^{-2}\right] + \\ c_{2|3}p_3 \times \phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\sigma_3^{-1}\frac{\partial \theta_2}{\partial \sigma_3} - (\theta_2-\mu_3)\sigma_3^{-2}\right] - \\ c_{2|3}p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\sigma_3^{-1}\frac{\partial \theta_1}{\partial \sigma_3} - (\theta_1-\mu_3)\sigma_3^{-2}\right] \end{bmatrix}
\end{aligned}
$$

and continuing to simplify:

$$
\begin{aligned}
\left(\frac{\partial BC}{\partial \sigma_3}\right) &= \begin{bmatrix}
\sigma_1^{-1} c_{2|1} p_1 \times \phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_2}{\partial \sigma_3}\right] - \\
\sigma_1^{-1} c_{2|1} p_1 \times \phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_1}{\partial \sigma_3}\right] + \\
\sigma_1^{-1} c_{3|1} p_1 \times \phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\left[-\frac{\partial \theta_2}{\partial \sigma_3}\right] + \\
\sigma_2^{-1} c_{1|2} p_2 \times \phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial \theta_1}{\partial \sigma_3}\right] + \\
\sigma_2^{-1} c_{3|2} p_2 \times \phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[-\frac{\partial \theta_2}{\partial \sigma_3}\right] + \\
\sigma_3^{-1} c_{1|3} p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_1}{\partial \sigma_3} - \left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right] + \\
\sigma_3^{-1} c_{2|3} p_3 \times \phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_2}{\partial \sigma_3} - \left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\right] - \\
\sigma_3^{-1} c_{2|3} p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_1}{\partial \sigma_3} - \left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right]
\end{bmatrix} \\
&= \begin{bmatrix}
\sigma_1^{-1} c_{2|1} p_1 \times \phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_2}{\partial \sigma_3}\right] - \\
\sigma_1^{-1} c_{2|1} p_1 \times \phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_1}{\partial \sigma_3}\right] + \\
\sigma_1^{-1} c_{3|1} p_1 \times \phi\left(\frac{\mu_1-\theta_2}{\sigma_1}\right)\left[-\frac{\partial \theta_2}{\partial \sigma_3}\right] + \\
\sigma_2^{-1} c_{1|2} p_2 \times \phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial \theta_1}{\partial \sigma_3}\right] + \\
\sigma_2^{-1} c_{3|2} p_2 \times \phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[-\frac{\partial \theta_2}{\partial \sigma_3}\right] + \\
\sigma_3^{-1} c_{1|3} p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_1}{\partial \sigma_3} - \left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right] + \\
\sigma_3^{-1} c_{2|3} p_3 \times \phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_2}{\partial \sigma_3} - \left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\right] - \\
\sigma_3^{-1} c_{2|3} p_3 \times \phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_1}{\partial \sigma_3} - \left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right]
\end{bmatrix} \\
&= \begin{bmatrix}
\sigma_1^{-1}\left[\phi\left(\frac{\theta_2-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_2}{\partial \sigma_3}\right](c_{2|1}p_1 - c_{3|1}p_1) - c_{2|1}p_1\phi\left(\frac{\theta_1-\mu_1}{\sigma_1}\right)\left[\frac{\partial \theta_1}{\partial \sigma_3}\right]\right] + \\
\sigma_2^{-1}\left[c_{1|2}p_2\phi\left(\frac{\theta_1-\mu_2}{\sigma_2}\right)\left[\frac{\partial \theta_1}{\partial \sigma_3}\right] + c_{3|2}p_2\phi\left(\frac{\mu_2-\theta_2}{\sigma_2}\right)\left[-\frac{\partial \theta_2}{\partial \sigma_3}\right]\right] + \\
\sigma_3^{-1}\left[\phi\left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_1}{\partial \sigma_3} - \left(\frac{\theta_1-\mu_3}{\sigma_3}\right)\right](c_{1|3}p_3 - c_{2|3}p_3) + c_{2|3}p_3\phi\left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\left[\frac{\partial \theta_2}{\partial \sigma_3} - \left(\frac{\theta_2-\mu_3}{\sigma_3}\right)\right]\right]
\end{bmatrix}
\end{aligned}
$$

A.3 Partial derivatives of four-class Bayes Cost with respect to all distributional parameters.

$$\left(\frac{\partial BC}{\partial \mu_1}\right) = \sigma_1^{-1}\left[c_{2|1}\times\left[\phi\left(\frac{\theta_2^*-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_2^*}{\partial\mu_1}-1\right]-\phi\left(\frac{\theta_1^*-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_1^*}{\partial\mu_1}-1\right]\right]+c_{3|1}\times\left[\phi\left(\frac{\theta_3^*-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_3^*}{\partial\mu_1}-1\right]-\phi\left(\frac{\theta_2^*-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_3^*}{\partial\mu_1}-1\right]\right]+c_{4|1}\times\left[\phi\left(\frac{\mu_1-\theta_3^*}{\sigma_1}\right)\left[-\frac{\partial\theta_3^*}{\partial\mu_1}\right]\right]\right]$$
$$+\sigma_2^{-1}\left[c_{1|2}\times\left[\phi\left(\frac{\theta_1^*-\mu_2}{\sigma_2}\right)\frac{\partial\theta_1^*}{\partial\mu_1}\right]+c_{3|2}\times\left[\phi\left(\frac{\theta_3^*-\mu_2}{\sigma_2}\right)\frac{\partial\theta_3^*}{\partial\mu_1}-\phi\left(\frac{\theta_2^*-\mu_2}{\sigma_2}\right)\frac{\partial\theta_2^*}{\partial\mu_1}\right]+c_{4|2}\times\left[\phi\left(\frac{\mu_2^*-\theta_3}{\sigma_2}\right)(-1)\frac{\partial\theta_3^*}{\partial\mu_1}\right]\right]$$
$$+\sigma_3^{-1}\left[c_{1|3}\times\left[\phi\left(\frac{\theta_1^*-\mu_3}{\sigma_3}\right)\frac{\partial\theta_1^*}{\partial\mu_1}\right]+c_{2|3}\times\left[\phi\left(\frac{\theta_2^*-\mu_3}{\sigma_3}\right)\frac{\partial\theta_2^*}{\partial\mu_1}-\phi\left(\frac{\theta_1^*-\mu_3}{\sigma_3}\right)\frac{\partial\theta_1^*}{\partial\mu_1}\right]+c_{4|3}\times\left[\phi\left(\frac{\mu_3^*-\theta_3}{\sigma_3}\right)(-1)\frac{\partial\theta_3^*}{\partial\mu_1}\right]\right]$$
$$+\sigma_4^{-1}\left[c_{1|4}\times\left[\phi\left(\frac{\theta_1^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_1^*}{\partial\mu_1}\right]+c_{2|4}\times\left[\phi\left(\frac{\theta_2^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_2^*}{\partial\mu_1}-\phi\left(\frac{\theta_1^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_1^*}{\partial\mu_1}\right]+c_{3|4}\times\left[\phi\left(\frac{\theta_3^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_3^*}{\partial\mu_1}-\phi\left(\frac{\theta_2^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_2^*}{\partial\mu_1}\right]\right] \tag{A.4}$$

$$\left(\frac{\partial BC}{\partial \mu_2}\right) = \sigma_1^{-1}\left[c_{2|1}\times\left[\phi\left(\frac{\theta_2^*-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_2^*}{\partial\mu_2}\right]-\phi\left(\frac{\theta_1^*-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_1^*}{\partial\mu_2}\right]\right]+c_{3|1}\times\left[\phi\left(\frac{\theta_3^*-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_3^*}{\partial\mu_2}\right]-\phi\left(\frac{\theta_2^*-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_3^*}{\partial\mu_2}\right]\right]+c_{4|1}\times\left[\phi\left(\frac{\mu_1-\theta_3^*}{\sigma_1}\right)\left[(-1)\frac{\partial\theta_3^*}{\partial\mu_2}\right]\right]\right]$$
$$+\sigma_2^{-1}\left[c_{1|2}\times\left[\phi\left(\frac{\theta_1^*-\mu_2}{\sigma_2}\right)\left[\frac{\partial\theta_1^*}{\partial\mu_2}-1\right]\right]+c_{3|2}\times\left[\phi\left(\frac{\theta_3^*-\mu_2}{\sigma_2}\right)\left[\frac{\partial\theta_3^*}{\partial\mu_2}-1\right]-\phi\left(\frac{\theta_2^*-\mu_2}{\sigma_2}\right)\left[\frac{\partial\theta_2^*}{\partial\mu_2}-1\right]\right]+c_{4|2}\times\left[\phi\left(\frac{\mu_2^*-\theta_3}{\sigma_2}\right)\left[1-\frac{\partial\theta_3^*}{\partial\mu_2}\right]\right]\right]$$
$$+\sigma_3^{-1}\left[c_{1|3}\times\left[\phi\left(\frac{\theta_1^*-\mu_3}{\sigma_3}\right)\frac{\partial\theta_1^*}{\partial\mu_2}\right]+c_{2|3}\times\left[\phi\left(\frac{\theta_2^*-\mu_3}{\sigma_3}\right)\frac{\partial\theta_2^*}{\partial\mu_2}-\phi\left(\frac{\theta_1^*-\mu_3}{\sigma_3}\right)\frac{\partial\theta_1^*}{\partial\mu_2}\right]+c_{4|3}\times\left[\phi\left(\frac{\mu_3^*-\theta_3}{\sigma_3}\right)(-1)\frac{\partial\theta_3^*}{\partial\mu_2}\right]\right]$$
$$+\sigma_4^{-1}\left[c_{1|4}\times\left[\phi\left(\frac{\theta_1^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_1^*}{\partial\mu_2}\right]+c_{2|4}\times\left[\phi\left(\frac{\theta_2^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_2^*}{\partial\mu_2}-\phi\left(\frac{\theta_1^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_1^*}{\partial\mu_2}\right]+c_{3|4}\times\left[\phi\left(\frac{\theta_3^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_3^*}{\partial\mu_2}-\phi\left(\frac{\theta_2^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_2^*}{\partial\mu_2}\right]\right] \tag{A.5}$$

$$\left(\frac{\partial BC}{\partial \mu_3}\right) = \sigma_1^{-1}\left[c_{2|1}\times\left[\phi\left(\frac{\theta_2^*-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_2^*}{\partial\mu_3}\right]-\phi\left(\frac{\theta_1^*-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_1^*}{\partial\mu_3}\right]\right]+c_{3|1}\times\left[\phi\left(\frac{\theta_3^*-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_3^*}{\partial\mu_3}\right]-\phi\left(\frac{\theta_2^*-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_3^*}{\partial\mu_3}\right]\right]+c_{4|1}\times\left[\phi\left(\frac{\mu_1-\theta_3^*}{\sigma_1}\right)\left[(-1)\frac{\partial\theta_3^*}{\partial\mu_3}\right]\right]\right]$$
$$+\sigma_2^{-1}\left[c_{1|2}\times\left[\phi\left(\frac{\theta_1^*-\mu_2}{\sigma_2}\right)\frac{\partial\theta_1^*}{\partial\mu_3}\right]+c_{3|2}\times\left[\phi\left(\frac{\theta_3^*-\mu_2}{\sigma_2}\right)\frac{\partial\theta_3^*}{\partial\mu_3}-\phi\left(\frac{\theta_2^*-\mu_2}{\sigma_2}\right)\frac{\partial\theta_2^*}{\partial\mu_3}\right]+c_{4|2}\times\left[\phi\left(\frac{\mu_2^*-\theta_3}{\sigma_2}\right)(-1)\frac{\partial\theta_3^*}{\partial\mu_3}\right]\right]$$
$$+\sigma_3^{-1}\left[c_{1|3}\times\left[\phi\left(\frac{\theta_1^*-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_1^*}{\partial\mu_3}-1\right]\right]+c_{2|3}\times\left[\phi\left(\frac{\theta_2^*-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_2^*}{\partial\mu_3}-1\right]-\phi\left(\frac{\theta_1^*-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_1^*}{\partial\mu_3}-1\right]\right]+c_{4|3}\times\left[\phi\left(\frac{\mu_3^*-\theta_3}{\sigma_3}\right)\left[-\frac{\partial\theta_3^*}{\partial\mu_3}1\right]\right]\right]$$
$$+\sigma_4^{-1}\left[c_{1|4}\times\left[\phi\left(\frac{\theta_1^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_1^*}{\partial\mu_3}\right]+c_{2|4}\times\left[\phi\left(\frac{\theta_2^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_2^*}{\partial\mu_3}-\phi\left(\frac{\theta_1^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_1^*}{\partial\mu_3}\right]+c_{3|4}\times\left[\phi\left(\frac{\theta_3^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_3^*}{\partial\mu_3}-\phi\left(\frac{\theta_2^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_2^*}{\partial\mu_3}\right]\right] \tag{A.6}$$

$$\left(\frac{\partial BC}{\partial \mu_4}\right) = \sigma_1^{-1}\left[c_{2|1}\times\left[\phi\left(\frac{\theta_2^*-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_2^*}{\partial\mu_4}\right]-\phi\left(\frac{\theta_1^*-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_1^*}{\partial\mu_4}\right]\right]+c_{3|1}\times\left[\phi\left(\frac{\theta_3^*-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_3^*}{\partial\mu_4}\right]-\phi\left(\frac{\theta_2^*-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_3^*}{\partial\mu_4}\right]\right]+c_{4|1}\times\left[\phi\left(\frac{\mu_1-\theta_3^*}{\sigma_1}\right)\left[(-1)\frac{\partial\theta_3^*}{\partial\mu_4}\right]\right]\right]$$
$$+\sigma_2^{-1}\left[c_{1|2}\times\left[\phi\left(\frac{\theta_1^*-\mu_2}{\sigma_2}\right)\frac{\partial\theta_1^*}{\partial\mu_4}\right]+c_{3|2}\times\left[\phi\left(\frac{\theta_3^*-\mu_2}{\sigma_2}\right)\frac{\partial\theta_3^*}{\partial\mu_4}-\phi\left(\frac{\theta_2^*-\mu_2}{\sigma_2}\right)\frac{\partial\theta_2^*}{\partial\mu_4}\right]+c_{4|2}\times\left[\phi\left(\frac{\mu_2^*-\theta_3}{\sigma_2}\right)(-1)\frac{\partial\theta_3^*}{\partial\mu_4}\right]\right]$$
$$+\sigma_3^{-1}\left[c_{1|3}\times\left[\phi\left(\frac{\theta_1^*-\mu_3}{\sigma_3}\right)\frac{\partial\theta_1^*}{\partial\mu_4}\right]+c_{2|3}\times\left[\phi\left(\frac{\theta_2^*-\mu_3}{\sigma_3}\right)\frac{\partial\theta_2^*}{\partial\mu_4}-\phi\left(\frac{\theta_1^*-\mu_3}{\sigma_3}\right)\frac{\partial\theta_1^*}{\partial\mu_4}\right]+c_{4|3}\times\left[\phi\left(\frac{\mu_3^*-\theta_3}{\sigma_3}\right)(-1)\frac{\partial\theta_3^*}{\partial\mu_4}\right]\right]$$
$$+\sigma_4^{-1}\left[c_{1|4}\times\left[\phi\left(\frac{\theta_1^*-\mu_4}{\sigma_4}\right)\left[\frac{\partial\theta_1^*}{\partial\mu_4}-1\right]\right]+c_{2|4}\times\left[\phi\left(\frac{\theta_2^*-\mu_4}{\sigma_4}\right)\left[\frac{\partial\theta_2^*}{\partial\mu_4}-1\right]-\phi\left(\frac{\theta_1^*-\mu_4}{\sigma_4}\right)\left[\frac{\partial\theta_1^*}{\partial\mu_4}-1\right]\right]+c_{3|4}\times\left[\phi\left(\frac{\theta_3^*-\mu_4}{\sigma_4}\right)\left[\frac{\partial\theta_3^*}{\partial\mu_4}-1\right]-\phi\left(\frac{\theta_2^*-\mu_4}{\sigma_4}\right)\left[\frac{\partial\theta_2^*}{\partial\mu_4}-1\right]\right]\right] \tag{A.7}$$

$$\left(\frac{\partial BC}{\partial \sigma_1}\right) = \sigma_1^{-1}\left[\begin{array}{l} c_{2|1}\times\left[\phi\left(\frac{\theta_2^*-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_2^*}{\partial\sigma_1}+\left(\frac{\theta_2^*-\mu_1}{\sigma_1}\right)\right]-\phi\left(\frac{\theta_1^*-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_1^*}{\partial\sigma_1}+\left(\frac{\theta_1^*-\mu_1}{\sigma_1}\right)\right]\right] \\ +c_{3|1}\times\left[\phi\left(\frac{\theta_3^*-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_3^*}{\partial\sigma_1}+\left(\frac{\theta_3^*-\mu_1}{\sigma_1}\right)\right]-\phi\left(\frac{\theta_2^*-\mu_1}{\sigma_1}\right)\left[\frac{\partial\theta_2^*}{\partial\sigma_1}+\left(\frac{\theta_2^*-\mu_1}{\sigma_1}\right)\right]\right]+c_{4|1}\times\left[\phi\left(\frac{\mu_1-\theta_3^*}{\sigma_1}\right)\left[(-1)\frac{\partial\theta_3^*}{\partial\sigma_1}+\left(\frac{\mu_1-\theta_3^*}{\sigma_1}\right)\right]\right] \end{array}\right]$$
$$+\sigma_2^{-1}\left[c_{1|2}\times\left[\phi\left(\frac{\theta_1^*-\mu_2}{\sigma_2}\right)\frac{\partial\theta_1^*}{\partial\sigma_1}\right]+c_{3|2}\times\left[\phi\left(\frac{\theta_3^*-\mu_2}{\sigma_2}\right)\frac{\partial\theta_3^*}{\partial\sigma_1}-\phi\left(\frac{\theta_2^*-\mu_2}{\sigma_2}\right)\frac{\partial\theta_3^*}{\partial\sigma_1}\right]+c_{4|2}\times\left[\phi\left(\frac{\mu_2^*-\theta_3}{\sigma_2}\right)(-1)\frac{\partial\theta_3^*}{\partial\sigma_1}\right]\right]$$
$$+\sigma_3^{-1}\left[c_{1|3}\times\left[\phi\left(\frac{\theta_1^*-\mu_3}{\sigma_3}\right)\frac{\partial\theta_1^*}{\partial\sigma_1}\right]+c_{2|3}\times\left[\phi\left(\frac{\theta_2^*-\mu_3}{\sigma_3}\right)\frac{\partial\theta_2^*}{\partial\sigma_1}-\phi\left(\frac{\theta_1^*-\mu_3}{\sigma_3}\right)\frac{\partial\theta_1^*}{\partial\sigma_1}\right]+c_{4|3}\times\left[\phi\left(\frac{\mu_3^*-\theta_3}{\sigma_3}\right)(-1)\frac{\partial\theta_3^*}{\partial\sigma_1}\right]\right]$$
$$+\sigma_4^{-1}\left[c_{1|4}\times\left[\phi\left(\frac{\theta_1^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_1^*}{\partial\sigma_1}\right]+c_{2|4}\times\left[\phi\left(\frac{\theta_2^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_2^*}{\partial\sigma_1}-\phi\left(\frac{\theta_1^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_1^*}{\partial\sigma_1}\right]+c_{3|4}\times\left[\phi\left(\frac{\theta_3^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_3^*}{\partial\sigma_1}-\phi\left(\frac{\theta_2^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_2^*}{\partial\sigma_1}\right]\right] \tag{A.8}$$

$$\left(\frac{\partial BC}{\partial \sigma_2}\right) = \sigma_1^{-1}\left[c_{2|1}\times\left[\phi\left(\frac{\theta_2^*-\mu_1}{\sigma_1}\right)\frac{\partial\theta_2^*}{\partial\sigma_2}-\phi\left(\frac{\theta_1^*-\mu_1}{\sigma_1}\right)\frac{\partial\theta_1^*}{\partial\sigma_2}\right]+c_{3|1}\times\left[\phi\left(\frac{\theta_3^*-\mu_1}{\sigma_1}\right)\frac{\partial\theta_3^*}{\partial\sigma_2}-\phi\left(\frac{\theta_2^*-\mu_1}{\sigma_1}\right)\frac{\partial\theta_2^*}{\partial\sigma_2}\right]+c_{4|1}\times\left[\phi\left(\frac{\mu_1-\theta_3^*}{\sigma_1}\right)(-1)\frac{\partial\theta_3^*}{\partial\sigma_2}\right]\right]$$
$$+\sigma_2^{-1}\left[\begin{array}{l} c_{1|2}\times\left[\phi\left(\frac{\theta_1^*-\mu_2}{\sigma_2}\right)\left[\frac{\partial\theta_1^*}{\partial\sigma_2}+\left(\frac{\theta_1^*-\mu_2}{\sigma_2}\right)\right]+c_{3|2}\times\left[\phi\left(\frac{\theta_3^*-\mu_2}{\sigma_2}\right)\left[\frac{\partial\theta_3^*}{\partial\sigma_2}+\left(\frac{\theta_3^*-\mu_2}{\sigma_2}\right)\right]-\phi\left(\frac{\theta_2^*-\mu_2}{\sigma_2}\right)\left[\frac{\partial\theta_2^*}{\partial\sigma_2}+\left(\frac{\theta_2^*-\mu_2}{\sigma_2}\right)\right]\right] \\ +c_{4|2}\times\left[\phi\left(\frac{\mu_2^*-\theta_3}{\sigma_2}\right)\left[(-1)\frac{\partial\theta_3^*}{\partial\sigma_2}+\left(\frac{\mu_2^*-\theta_3}{\sigma_2}\right)\right]\right] \end{array}\right]$$
$$+\sigma_3^{-1}\left[c_{1|3}\times\left[\phi\left(\frac{\theta_1^*-\mu_3}{\sigma_3}\right)\frac{\partial\theta_1^*}{\partial\sigma_2}\right]+c_{2|3}\times\left[\phi\left(\frac{\theta_2^*-\mu_3}{\sigma_3}\right)\frac{\partial\theta_2^*}{\partial\sigma_2}-\phi\left(\frac{\theta_1^*-\mu_3}{\sigma_3}\right)\frac{\partial\theta_1^*}{\partial\sigma_2}\right]+c_{4|3}\times\left[\phi\left(\frac{\mu_3^*-\theta_3}{\sigma_3}\right)(-1)\frac{\partial\theta_3^*}{\partial\sigma_2}\right]\right]$$
$$+\sigma_4^{-1}\left[c_{1|4}\times\left[\phi\left(\frac{\theta_1^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_1^*}{\partial\sigma_2}\right]+c_{2|4}\times\left[\phi\left(\frac{\theta_2^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_2^*}{\partial\sigma_2}-\phi\left(\frac{\theta_1^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_1^*}{\partial\sigma_2}\right]+c_{3|4}\times\left[\phi\left(\frac{\theta_3^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_3^*}{\partial\sigma_2}-\phi\left(\frac{\theta_2^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_2^*}{\partial\sigma_2}\right]\right] \tag{A.9}$$

$$\left(\frac{\partial BC}{\partial \sigma_3}\right) = \sigma_1^{-1}\left[c_{2|1}\times\left[\phi\left(\frac{\theta_2^*-\mu_1}{\sigma_1}\right)\frac{\partial\theta_2^*}{\partial\sigma_3}-\phi\left(\frac{\theta_1^*-\mu_1}{\sigma_1}\right)\frac{\partial\theta_1^*}{\partial\sigma_3}\right]+c_{3|1}\times\left[\phi\left(\frac{\theta_3^*-\mu_1}{\sigma_1}\right)\frac{\partial\theta_3^*}{\partial\sigma_3}-\phi\left(\frac{\theta_2^*-\mu_1}{\sigma_1}\right)\frac{\partial\theta_2^*}{\partial\sigma_3}\right]+c_{4|1}\times\left[\phi\left(\frac{\mu_1-\theta_3^*}{\sigma_1}\right)(-1)\frac{\partial\theta_3^*}{\partial\sigma_3}\right]\right]$$
$$+\sigma_2^{-1}\left[c_{1|2}\times\left[\phi\left(\frac{\theta_1^*-\mu_2}{\sigma_2}\right)\frac{\partial\theta_1^*}{\partial\sigma_3}\right]+c_{3|2}\times\left[\phi\left(\frac{\theta_3^*-\mu_2}{\sigma_2}\right)\frac{\partial\theta_3^*}{\partial\sigma_3}-\phi\left(\frac{\theta_2^*-\mu_2}{\sigma_2}\right)\frac{\partial\theta_2^*}{\partial\sigma_3}\right]+c_{4|2}\times\left[\phi\left(\frac{\mu_2^*-\theta_3}{\sigma_2}\right)(-1)\frac{\partial\theta_3^*}{\partial\sigma_3}\right]\right]$$
$$+\sigma_3^{-1}\left[\begin{array}{l} c_{1|3}\times\left[\phi\left(\frac{\theta_1^*-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_1^*}{\partial\sigma_3}+\left(\frac{\theta_1^*-\mu_3}{\sigma_3}\right)\right]\right] \\ +c_{2|3}\times\left[\phi\left(\frac{\theta_2^*-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_2^*}{\partial\sigma_3}+\left(\frac{\theta_2^*-\mu_3}{\sigma_3}\right)\right]-\phi\left(\frac{\theta_1^*-\mu_3}{\sigma_3}\right)\left[\frac{\partial\theta_1^*}{\partial\sigma_3}+\left(\frac{\theta_1^*-\mu_3}{\sigma_3}\right)\right]\right]+c_{4|3}\times\left[\phi\left(\frac{\mu_3^*-\theta_3}{\sigma_3}\right)\left[(-1)\frac{\partial\theta_3^*}{\partial\sigma_3}+\left(\frac{\mu_3^*-\theta_3}{\sigma_3}\right)\right]\right]\end{array}\right]$$
$$+\sigma_4^{-1}\left[c_{1|4}\times\left[\phi\left(\frac{\theta_1^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_1^*}{\partial\sigma_3}\right]+c_{2|4}\times\left[\phi\left(\frac{\theta_2^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_2^*}{\partial\sigma_3}-\phi\left(\frac{\theta_1^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_1^*}{\partial\sigma_3}\right]+c_{3|4}\times\left[\phi\left(\frac{\theta_3^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_3^*}{\partial\sigma_3}-\phi\left(\frac{\theta_2^*-\mu_4}{\sigma_4}\right)\frac{\partial\theta_2^*}{\partial\sigma_3}\right]\right] \quad \text{(A.10)}$$

$$\left(\frac{\partial BC}{\partial \sigma_4}\right) = \sigma_1^{-1}\left[c_{2|1}\times\left[\phi\left(\frac{\theta_2^*-\mu_1}{\sigma_1}\right)\frac{\partial\theta_2^*}{\partial\sigma_4}-\phi\left(\frac{\theta_1^*-\mu_1}{\sigma_1}\right)\frac{\partial\theta_1^*}{\partial\sigma_4}\right]+c_{3|1}\times\left[\phi\left(\frac{\theta_3^*-\mu_1}{\sigma_1}\right)\frac{\partial\theta_3^*}{\partial\sigma_4}-\phi\left(\frac{\theta_2^*-\mu_1}{\sigma_1}\right)\frac{\partial\theta_2^*}{\partial\sigma_4}\right]+c_{4|1}\times\left[\phi\left(\frac{\mu_1-\theta_3^*}{\sigma_1}\right)(-1)\frac{\partial\theta_3^*}{\partial\sigma_4}\right]\right]$$
$$+\sigma_2^{-1}\left[c_{1|2}\times\left[\phi\left(\frac{\theta_1^*-\mu_2}{\sigma_2}\right)\frac{\partial\theta_1^*}{\partial\sigma_4}\right]+c_{3|2}\times\left[\phi\left(\frac{\theta_3^*-\mu_2}{\sigma_2}\right)\frac{\partial\theta_3^*}{\partial\sigma_4}-\phi\left(\frac{\theta_2^*-\mu_2}{\sigma_2}\right)\frac{\partial\theta_3^*}{\partial\sigma_4}\right]+c_{4|2}\times\left[\phi\left(\frac{\mu_2^*-\theta_3}{\sigma_2}\right)(-1)\frac{\partial\theta_3^*}{\partial\sigma_4}\right]\right]$$
$$+\sigma_3^{-1}\left[c_{1|3}\times\left[\phi\left(\frac{\theta_1^*-\mu_3}{\sigma_3}\right)\frac{\partial\theta_1^*}{\partial\sigma_4}\right]+c_{2|3}\times\left[\phi\left(\frac{\theta_2^*-\mu_3}{\sigma_4}\right)\frac{\partial\theta_2^*}{\partial\sigma_2}-\phi\left(\frac{\theta_1^*-\mu_3}{\sigma_4}\right)\frac{\partial\theta_1^*}{\partial\sigma_2}\right]+c_{4|3}\times\left[\phi\left(\frac{\mu_3^*-\theta_3}{\sigma_4}\right)(-1)\frac{\partial\theta_3^*}{\partial\sigma_2}\right]\right]$$
$$+\sigma_4^{-1}\left[\begin{array}{l} c_{1|4}\times\left[\phi\left(\frac{\theta_1^*-\mu_4}{\sigma_4}\right)\left[\frac{\partial\theta_1^*}{\partial\sigma_4}+\left(\frac{\theta_1^*-\mu_4}{\sigma_4}\right)\right]\right]+c_{2|4}\times\left[\phi\left(\frac{\theta_2^*-\mu_4}{\sigma_4}\right)\left[\frac{\partial\theta_2^*}{\partial\sigma_4}+\left(\frac{\theta_2^*-\mu_4}{\sigma_4}\right)\right]-\phi\left(\frac{\theta_1^*-\mu_4}{\sigma_4}\right)\left[\frac{\partial\theta_1^*}{\partial\sigma_4}+\left(\frac{\theta_1^*-\mu_4}{\sigma_4}\right)\right]\right] \\ +c_{3|4}\times\left[\phi\left(\frac{\theta_3^*-\mu_4}{\sigma_4}\right)\left[\frac{\partial\theta_3^*}{\partial\sigma_4}+\left(\frac{\theta_3^*-\mu_4}{\sigma_4}\right)\right]-\phi\left(\frac{\theta_2^*-\mu_4}{\sigma_4}\right)\left[\frac{\partial\theta_2^*}{\partial\sigma_4}+\left(\frac{\theta_2^*-\mu_4}{\sigma_4}\right)\right]\right]\end{array}\right] \quad \text{(A.11)}$$

A.4 Wald and Log Wald CI for Bayes Cost

The Wald method for constructing CIs is common and easily applied for large sample sizes, though may not perform as well as other methods. Developed with the large sample normality of maximum likelihood estimators (MLEs), the statistic

$$z = \frac{\widehat{\theta}-\theta_0}{S_{\widehat{\theta}}} \quad \text{(A.12)}$$

is approximately standard normal when $\theta = \theta_0$ [3, p. 11],[70]. The $(1-\alpha)100\%$ Wald CI is then found as

$$\widehat{\theta} \pm Z_{1-\frac{\alpha}{2}} \times S_{\widehat{\theta}} \quad \text{(A.13)}$$

Although the Wald CI is easy to implement, it performs poorly for binomial probabilities with respect to coverage [2]. Despite this, the Wald CI is considered for BC as it is easily computed and a good place to start for baseline comparison of newly developed methods, and may perform better for the sum of binomial/multinomial probabilities (i.e. BC) rather than for binomial probabilities directly.

A.4.1 Bayes Cost for two-class classification system.

Consider a two-class classification system with results tabulated in a contingency table as in Table 2.3. Class one has $n_1 = X_{1|1}+X_{2|1}$ observations and class two has $n_2 = X_{1|2}+X_{2|2}$ observations (with n_1 and n_2 fixed). The outcomes from each class are mutually exclusive and independently

distributed, and for each observation in a class, the classification system labels each observation as only one of the two possible outcomes. No distributional assumptions on the feature or features used for classification are made. In [76], J is defined as the maximum of the sum of correct classification rates minus one, which can be written as

$$J = \max_{\theta \in \Theta} \left[\frac{X_{1|1}}{n_1} + \frac{X_{2|2}}{n_2} - 1 \right] \tag{A.14}$$

where $X_{1|1}$ and $X_{2|2}$ are the random variables representing the number of observations correctly classified for a vector of thresholds $\boldsymbol{\theta} \in \Theta$. Bayes Cost, which is defined to minimize the misclassification rates instead of maximizing the correct classification rates, can be used similarly. In the nonparametric framework, BC (with equal cost and prevalence multipliers, assumed to equal one) may be written

$$BC = \min_{\boldsymbol{\theta} \in \Theta} \left[\frac{X_{2|1}}{n_1} + \frac{X_{1|2}}{n_2} \right] \tag{A.15}$$

where $X_{2|1}$ and $X_{1|2}$ are the random variables representing the misclassified observations for a $\boldsymbol{\theta} \in \Theta$. The expected value and variance of BC defined in Equation A.15 is determined using properties of the binomial distribution.

$$\begin{aligned} E(BC) &= E\left(\frac{X_{2|1}}{n_1} + \frac{X_{1|2}}{n_2} \right) \\ &= \frac{1}{n_1} E(X_{2|1}) + \frac{1}{n_2} E(X_{1|2}) \\ &= \frac{P_{2|1}(\boldsymbol{\theta}) \times n_1}{n_1} + \frac{P_{1|2}(\boldsymbol{\theta}) \times n_2}{n_2} \\ &= P_{2|1}(\boldsymbol{\theta}) + P_{1|2}(\boldsymbol{\theta}) \end{aligned} \tag{A.16}$$

$$\begin{aligned} Var(BC) &= Var\left(\frac{X_{2|1}}{n_1} + \frac{X_{1|2}}{n_2} \right) \\ &= \frac{1}{n_1^2} Var(X_{2|1}) + \frac{1}{n_2^2} Var(X_{1|2}) \\ &= \frac{P_{2|1}(\boldsymbol{\theta}) \times P_{1|1}(\boldsymbol{\theta}) \times n_1}{n_1^2} + \frac{P_{1|2}(\boldsymbol{\theta}) \times P_{2|2}(\boldsymbol{\theta}) \times n_2}{n_2^2} \\ &= \frac{P_{2|1}(\boldsymbol{\theta}) \times P_{1|1}(\boldsymbol{\theta})}{n_1} + \frac{P_{1|2}(\boldsymbol{\theta}) \times P_{2|2}(\boldsymbol{\theta})}{n_2} \end{aligned} \tag{A.17}$$

where $P_{i|j}(\boldsymbol{\theta})$ is the true probability of classifying class j as class i for a given $\boldsymbol{\theta} \in \Theta$ and n_j is the total number of observations sampled from the j^{th} class. Using the MLEs for $P_{i|j}(\boldsymbol{\theta})$ (from the binomial distribution presented in Section 2.7.1), BC and the variance of $\widehat{BC}$ are estimated

$$\widehat{BC} = \frac{x_{2|1}}{n_1} + \frac{x_{1|2}}{n_2} \tag{A.18}$$

$$Var(\widehat{BC}) = \frac{x_{1|1}x_{2|1}}{(n_1)^3} + \frac{x_{1|2}x_{2|2}}{(n_2)^3} \tag{A.19}$$

For greater utility, BC may be defined with prevalences on classes and different costs on misclassification errors [58, 65] such that

$$BC = \min_{\boldsymbol{\theta}\in\Theta} \left[c_{2|1}p_1 \frac{X_{2|1}}{n_1} + c_{1|2}p_2 \frac{X_{1|2}}{n_2} \right] \tag{A.20}$$

where $c_{i|j}$ is the fixed cost associated with misclassifying class j as class i and p_j is the fixed prevalence for the j^{th} class. The expected value and variance of BC defined in Equation A.20 is

$$E[BC] = p_1 c_{2|1} P_{2|1}(\boldsymbol{\theta}) + p_2 c_{1|2} P_{1|2}(\boldsymbol{\theta}) \tag{A.21}$$

$$Var[BC] = (p_1 c_{2|1})^2 \frac{P_{2|1}(\boldsymbol{\theta}) \times P_{1|1}(\boldsymbol{\theta})}{n_1} + (p_2 c_{1|2})^2 \frac{P_{1|2}(\boldsymbol{\theta}) \times P_{2|2}(\boldsymbol{\theta})}{n_2} \tag{A.22}$$

Once again, using the MLEs for the binomial proportions $P_{i|j}(\boldsymbol{\theta})$, $\widehat{BC}$ and the variance of $\widehat{BC}$ are

$$\widehat{BC} = p_1 c_{2|1} \frac{x_{2|1}}{n_1} + p_2 c_{1|2} \frac{x_{1|2}}{n_2} \tag{A.23}$$

$$Var(\widehat{BC}) = (p_1 c_{2|1})^2 \frac{x_{1|1}x_{2|1}}{n_1^3} + (p_2 c_{1|2})^2 \frac{x_{1|2}x_{2|2}}{n_2^3} \tag{A.24}$$

A.4.2 Bayes Cost for a k-class classification system.

Consider a classification system with three or more classes where the diagnostic outcomes may be tabulated in a contingency table as in Table 2.4, for a given $\boldsymbol{\theta} \in \Theta$. Once again, no distributional assumptions on the feature or features used for classification are made. For the three-class example, the first class has $n_1 = x_{1|1} + x_{2|1} + x_{3|1}$ observations, the second class has $n_2 = x_{1|2} + x_{2|2} + x_{3|2}$, and the third class has $n_3 = x_{1|3} + x_{2|3} + x_{3|3}$ observations (where n_1, n_2, and n_3 are all assumed fixed). The outcomes from the classes are mutually exclusive with independent distributions and the classification system labels each observation as only one of the three (or k for k classes) possible outcomes. Therefore, the number of outcomes in each diagnostic state in a single class (or column

in the contingency table) are distributed multinomial (see Section 2.7.2). Similar to the two-class classification system, BC is defined with costs and class prevalence multipliers:

$$BC = \min_{\boldsymbol{\theta}\in\Theta} \left[\sum_{\substack{i=1 \\ i\neq j}}^{3} \sum_{j=1}^{3} c_{i|j} p_j \frac{X_{i|j}}{n_j} \right] \tag{A.25}$$

The expected value and variance of BC is determined directly from the properties of the multinomial distribution, taking into account the covariances between outcomes within the same class. Therefore,

$$E(BC) = \sum_{\substack{i=1 \\ i\neq j}}^{3} \sum_{j=1}^{3} c_{i|j} p_j P_{i|j}(\boldsymbol{\theta}) \tag{A.26}$$

and

$$Var(BC) = \sum_{j=1}^{3} \left(\sum_{\substack{i=1 \\ i\neq j}}^{3} \left(\frac{(c_{i|j} p_j)^2}{n_j} P_{i|j}(\boldsymbol{\theta}) \times (1 - P_{i|j}(\boldsymbol{\theta})) \right) - \frac{2p_j^2}{n_j} \prod_{\substack{i=1 \\ i\neq j}}^{3} c_{i|j} P_{i|j}(\boldsymbol{\theta}) \right) \tag{A.27}$$

The MLEs for the multinomial distribution are used to estimate BC and the variance of $\widehat{BC}$ as follows

$$\widehat{BC} = \sum_{\substack{i=1 \\ i\neq j}}^{3} \sum_{j=1}^{3} c_{i|j} p_j \frac{x_{i|j}}{n_j} \tag{A.28}$$

and

$$Var(\widehat{BC}) = \sum_{j=1}^{3} \left(\sum_{\substack{i=1 \\ i\neq j}}^{3} \left(\frac{(c_{i|j} p_j)^2 \times x_{i|j}}{n_j^2} \times (1 - \frac{x_{i|j}}{n_j}) \right) - \frac{2p_j^2}{n_j} \prod_{\substack{i=1 \\ i\neq j}}^{3} c_{i|j} \frac{x_{i|j}}{n_j} \right) \tag{A.29}$$

Equation A.25 can be generalized for k classes as [58]

$$BC = \min_{\boldsymbol{\theta}\in\Theta} \left[\sum_{\substack{i=1 \\ i\neq j}}^{k} \sum_{j=1}^{k} c_{i|j} p_j \frac{X_{i|j}}{n_j} \right] \tag{A.30}$$

Further, $\widehat{BC}$ and $Var(\widehat{BC})$ for any k-class BC is found similar to Equations A.28 and A.29 using the mean, variance, and covariance of multinomial random variables. Although an equivalence between the optimal threshold for the two-class BC and the GYI optimal threshold exists (see Theorem 3, Section 2.5.2), for $k \geq 3$ classes this equivalence of optimal thresholds does not universally hold, specifically when the costs of misclassification within a single class or between classes are not equal [58].

A.4.3 Wald and Log Wald Confidence Intervals.

A $(1-\alpha)100\%$ Wald CI for the k-class BC ($k = 2, 3, \dots$) is

$$\widehat{BC} \pm Z_{(1-\frac{\alpha}{2})} \sqrt{Var(\widehat{BC})} \tag{A.31}$$

where the $\widehat{BC}$ and $Var(\widehat{BC})$ are found nonparametrically as in Sections A.4.1 and A.4.2. Since BC is bounded above zero, a CI around the natural logarithm of BC is also considered in order to assure that the CI greater than zero [41, p.163]. The $(1-\alpha)100\%$ Wald CI around the log of BC is

$$\log(\widehat{BC}) \pm z_{\alpha/2} \times Var(\log(\widehat{BC})) \tag{A.32}$$

Then the $(1-\alpha)100\%$ log Wald CI for BC is

$$BC \times \exp\left[\pm z_{\alpha/2} \times Var(\log(\widehat{BC}))\right] \tag{A.33}$$

where the delta method is used to approximate $Var(\log(\widehat{BC}))$ as

$$Var(\log(\widehat{BC})) \approx \left(\frac{\partial \log(\widehat{BC})}{\partial \widehat{BC}}\right)^2 Var(\widehat{BC}) = \frac{1}{\widehat{BC}^2} Var(\widehat{BC}) \tag{A.34}$$

A.5 Regularity Conditions

Regularity conditions required for Theorem 8 are given in [12, p.516], listed below. These conditions are assumed for the normal and multinomial distributions, which are exponential family distributions.

(A1) We observe $X_1, \ldots, X_n$ where $X_i \sim f(x \mid \theta)$ are iid.
(A2) The parameter is *identifiable*; that is, if $\theta \neq \theta'$, then $f(x \mid \theta) \neq f(x \mid \theta')$.
(A3) The densities $f(x \mid \theta)$ have common support, and $f(x \mid \theta)$ is differentiable in θ .
(A4) The parameter space Ω contains an open set ω of which the true parameter value θ_0 is an interior point.
(A5) For every $x \in \mathcal{X}$, the density $f(x \mid \theta)$ is three times differentiable with respect to θ , the third derivative is continuous in θ , and $\int f(x \mid \theta)dx$ can be differentiated three times under the integral sign.
(A6) For any $\theta_0 \in \Omega$, there exists a positive number c and a function $M(x)$ (both of which may depend on θ_0) such that $\left|\frac{\partial^3}{\partial\theta^3} \log f(x \mid \theta)\right| \leq M(x)$ for all $x \in \mathcal{X}$, $\theta_0 - c < \theta < \theta_0 + c$, with $E_{\theta_0}[M(X)] < \infty$.

Appendix B: Additional Tables

B.1 Parametric Confidence Interval Simulation Tables

Table B.1: Coverage probability and length for parametric 95% CIs around *BC* under equal costs and three classes with a normally distributed feature.

			Delta		GCI		BCa		BP		AN	
	n_j	BC_3	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len
Normal	10	1.23	92.60	0.70	96.13	0.67	88.20	0.63	90.26	0.68	83.82	0.68
$\sigma_3 = 1$		0.91	91.98	0.69	96.03	0.68	89.78	0.63	89.36	0.66	85.54	0.66
		0.63	91.78	0.65	96.07	0.65	89.88	0.58	87.52	0.58	85.16	0.58
		0.42	91.02	0.57	95.83	0.60	89.42	0.51	85.24	0.48	84.66	0.47
		0.27	90.04	0.46	95.47	0.52	89.36	0.41	83.12	0.37	84.98	0.37
	50	1.23	94.80	0.32	95.63	0.32	94.48	0.32	94.42	0.32	93.24	0.32
		0.91	94.78	0.32	95.70	0.32	94.46	0.31	94.28	0.31	93.52	0.31
		0.63	94.44	0.30	95.77	0.30	94.14	0.29	93.64	0.29	93.00	0.29
		0.42	94.02	0.26	95.43	0.26	94.12	0.26	93.08	0.25	93.20	0.25
		0.27	94.50	0.21	95.50	0.22	94.00	0.21	92.58	0.20	93.20	0.20
	100	1.23	94.56	0.23	95.47	0.23	94.30	0.23	94.44	0.23	93.48	0.23
		0.91	94.48	0.23	94.97	0.22	94.46	0.23	94.26	0.22	93.60	0.22
		0.63	95.12	0.21	94.93	0.21	94.22	0.21	93.84	0.21	93.80	0.21
		0.42	94.72	0.18	94.73	0.19	94.12	0.18	93.66	0.18	93.88	0.18
		0.27	94.70	0.15	94.57	0.15	94.12	0.15	93.74	0.15	93.60	0.15
	250	1.23	95.04	0.14	95.20	0.14	94.98	0.14	94.86	0.14	94.90	0.14
		0.91	94.82	0.14	95.30	0.14	94.86	0.14	94.58	0.14	94.88	0.14
		0.63	94.76	0.13	95.03	0.13	94.72	0.13	94.16	0.13	94.42	0.13
		0.42	94.70	0.12	94.77	0.12	94.54	0.12	94.24	0.12	94.62	0.12
		0.27	94.66	0.09	94.80	0.10	94.36	0.10	94.24	0.09	94.52	0.09
Normal	10	1.23	92.44	0.76	96.33	0.73	90.24	0.72	89.86	0.74	87.48	0.74
$\sigma_3 = 2$		0.91	91.66	0.74	96.57	0.73	91.44	0.70	89.52	0.70	86.74	0.70
		0.63	91.38	0.67	96.30	0.69	91.02	0.63	87.78	0.61	85.88	0.60
		0.42	90.74	0.58	96.27	0.62	90.08	0.53	85.96	0.50	85.10	0.49
		0.27	90.20	0.47	96.30	0.52	89.38	0.43	83.74	0.38	85.18	0.37
	50	1.23	94.54	0.35	95.43	0.35	94.74	0.35	94.38	0.35	93.80	0.35
		0.91	94.36	0.34	95.73	0.34	94.40	0.34	94.22	0.34	93.54	0.34
		0.63	94.04	0.31	95.73	0.31	94.26	0.31	93.82	0.30	93.62	0.30
		0.42	93.78	0.27	95.83	0.27	94.14	0.26	93.44	0.26	93.12	0.26
		0.27	93.40	0.21	95.57	0.22	93.92	0.21	92.90	0.20	93.34	0.20
	100	1.23	95.08	0.25	95.07	0.25	94.30	0.25	94.18	0.25	94.18	0.25
		0.91	95.34	0.24	95.20	0.24	94.38	0.24	94.24	0.24	94.02	0.24
		0.63	95.02	0.22	95.10	0.22	94.40	0.22	94.04	0.22	94.26	0.22
		0.42	94.86	0.19	95.20	0.19	94.30	0.19	94.02	0.19	94.10	0.19
		0.27	94.74	0.15	94.90	0.15	94.26	0.15	93.60	0.15	94.20	0.15

Continued on next page

Table B.1 – continued from previous page

			Delta		GCI		BCa		BP		AN	
	n_j	BC_3	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len
	250	1.23	95.12	0.16	95.17	0.16	94.94	0.16	94.88	0.16	94.86	0.16
		0.91	95.02	0.15	95.17	0.15	94.84	0.15	94.66	0.15	94.86	0.15
		0.63	94.68	0.14	95.07	0.14	94.78	0.14	94.72	0.14	94.68	0.14
		0.42	94.90	0.12	95.13	0.12	94.52	0.12	94.28	0.12	94.58	0.15
		0.27	94.68	0.10	95.07	0.10	94.46	0.10	94.52	0.09	94.60	0.10
Normal	10	1.23	92.28	0.77	96.43	0.75	91.84	0.75	90.40	0.76	89.28	0.76
$\sigma_3 = 4$		0.91	92.02	0.74	96.77	0.75	93.00	0.73	90.50	0.72	87.78	0.72
		0.63	91.56	0.67	96.90	0.70	92.26	0.65	88.94	0.62	86.68	0.62
		0.42	90.98	0.58	96.70	0.62	91.36	0.55	86.96	0.50	85.60	0.50
		0.27	90.50	0.47	96.80	0.52	89.88	0.44	84.56	0.38	85.62	0.38
	50	1.23	94.36	0.36	95.70	0.35	94.78	0.36	94.22	0.35	94.00	0.36
		0.91	94.12	0.34	95.53	0.34	94.86	0.34	94.52	0.34	93.90	0.34
		0.63	93.98	0.31	95.77	0.31	94.62	0.31	94.06	0.30	93.64	0.31
		0.42	93.60	0.27	95.90	0.27	94.24	0.27	93.50	0.26	93.28	0.26
		0.27	93.26	0.21	95.73	0.22	93.88	0.21	92.82	0.20	93.32	0.20
	100	1.23	95.28	0.25	95.10	0.25	94.30	0.25	94.16	0.25	93.98	0.25
		0.91	95.04	0.24	95.17	0.24	94.26	0.24	94.40	0.24	94.34	0.24
		0.63	94.86	0.22	95.10	0.22	94.40	0.22	94.20	0.22	94.14	0.22
		0.42	94.62	0.19	95.13	0.19	94.26	0.19	94.02	0.19	94.38	0.19
		0.27	94.62	0.15	95.30	0.15	94.04	0.15	93.90	0.15	94.16	0.15
	250	1.23	95.14	0.16	94.87	0.16	94.72	0.16	94.78	0.16	94.96	0.16
		0.91	95.06	0.15	95.07	0.15	94.90	0.15	94.86	0.15	94.98	0.15
		0.63	94.90	0.14	94.97	0.14	94.86	0.14	94.70	0.14	94.72	0.14
		0.42	94.76	0.12	94.93	0.12	94.80	0.12	94.58	0.12	94.90	0.12
		0.27	94.68	0.10	95.13	0.10	94.74	0.10	94.74	0.09	94.82	0.10

GCI - generalized confidence interval; BCa - bias corrected and accelerated; BP - basic percentile;
AN - asymptotic normal; Cov - coverage; Len - length

Table B.2: Coverage probability and length for 95% parametric CIs around *BC* under equal costs and three classes with a non-normally distributed feature.

			Delta		GCI		BCa		BP		AN	
	n_j	BC_3	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len
Gamma	10	1.23	88.68	0.66	92.20	0.64	89.96	0.73	90.20	0.72	79.44	0.72
		0.91	88.94	0.82	95.83	0.68	90.08	0.67	88.50	0.65	78.40	0.65
		0.63	89.10	0.61	96.30	0.61	90.84	0.66	88.92	0.61	83.78	0.61
		0.42	91.30	0.54	97.47	0.57	89.08	0.45	84.48	0.44	80.26	0.43
		0.27	94.22	0.46	98.60	0.49	89.38	0.36	85.74	0.34	86.64	0.33
	50	1.23	84.90	0.30	85.50	0.30	89.90	0.34	89.66	0.34	85.32	0.34
		0.91	84.60	0.31	87.67	0.31	87.38	0.31	86.08	0.31	79.34	0.31
		0.63	90.32	0.28	91.93	0.28	92.58	0.30	90.90	0.30	89.96	0.30
		0.42	93.98	0.25	95.93	0.25	91.80	0.23	90.50	0.22	88.56	0.22
		0.27	96.92	0.21	96.60	0.21	90.96	0.18	90.84	0.17	94.10	0.17
	100	1.23	81.26	0.21	81.73	0.21	88.64	0.25	87.94	0.25	83.36	0.25
		0.91	78.76	0.22	79.93	0.22	82.38	0.22	80.54	0.22	73.86	0.22
		0.63	89.86	0.20	90.80	0.20	93.18	0.22	91.74	0.22	90.76	0.22
		0.42	94.42	0.18	94.10	0.18	92.90	0.17	92.38	0.17	90.68	0.17
		0.27	96.12	0.15	94.70	0.15	89.18	0.13	90.72	0.13	94.04	0.13
	250	1.23	74.04	0.13	72.63	0.13	83.92	0.16	82.98	0.16	78.64	0.16
		0.91	63.38	0.14	62.83	0.14	70.50	0.14	67.62	0.14	62.14	0.14
		0.63	88.92	0.12	89.17	0.12	92.76	0.14	91.84	0.14	90.96	0.14
		0.42	94.32	0.11	95.07	0.11	93.70	0.11	93.14	0.11	91.78	0.11
		0.27	93.38	0.09	90.87	0.09	85.62	0.08	87.24	0.08	91.20	0.08
Gamma w/ Box-Cox	10	1.23	92.36	0.69	95.37	0.65	91.68	0.70	92.94	0.70	84.58	0.71
		0.91	91.52	0.69	95.67	0.67	90.68	0.66	89.66	0.68	85.50	0.67
		0.63	89.80	0.62	94.43	0.60	90.64	0.61	86.90	0.60	87.04	0.60
		0.42	91.58	0.57	94.03	0.59	89.40	0.50	85.56	0.48	85.78	0.48
		0.27	90.70	0.48	92.83	0.52	89.38	0.44	84.16	0.40	86.14	0.39
	50	1.23	94.32	0.32	95.43	0.40	94.20	0.31	94.24	0.31	91.72	0.31
		0.91	94.16	0.32	95.03	0.41	94.14	0.32	93.60	0.32	93.32	0.32
		0.63	92.98	0.29	94.00	0.37	94.02	0.30	93.54	0.30	94.14	0.30
		0.42	94.52	0.26	94.27	0.34	92.66	0.25	92.08	0.25	93.18	0.25
		0.27	92.96	0.22	91.53	0.29	90.84	0.23	91.32	0.22	93.66	0.22
	100	1.23	94.22	0.22	94.43	0.31	94.56	0.22	94.44	0.22	92.74	0.22
		0.91	93.86	0.23	94.53	0.32	94.14	0.23	93.96	0.23	93.76	0.23
		0.63	92.28	0.20	92.47	0.29	93.18	0.22	93.20	0.22	94.38	0.22
		0.42	94.82	0.18	94.60	0.26	93.22	0.18	93.28	0.18	94.48	0.18
		0.27	91.74	0.16	90.87	0.22	89.20	0.16	91.34	0.16	93.42	0.16
	250	1.23	93.92	0.14	94.40	0.22	94.18	0.14	94.30	0.14	92.98	0.14

Continued on next page

Table B.2 – continued from previous page

			Delta		GCI		BCa		BP		AN	
	n_j	BC_3	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len
		0.91	94.36	0.14	95.23	0.23	94.86	0.15	94.78	0.15	94.78	0.15
		0.63	92.40	0.13	92.80	0.20	93.80	0.14	94.10	0.14	94.74	0.14
		0.42	94.58	0.12	94.07	0.18	93.62	0.12	93.66	0.12	94.56	0.12
		0.27	85.76	0.10	88.13	0.16	82.72	0.10	85.56	0.10	88.52	0.10
Normal	10	1.23	87.84	2.39	93.20	0.72	86.60	0.68	87.28	0.70	79.32	0.70
Mixture		0.91	90.14	0.75	95.23	0.75	91.08	0.73	88.64	0.73	85.58	0.73
		0.63	87.54	0.67	94.63	0.69	89.90	0.66	85.40	0.62	81.64	0.62
		0.42	87.94	0.57	94.80	0.61	89.26	0.56	83.78	0.52	83.92	0.51
		0.27	84.08	0.44	93.67	0.49	89.36	0.47	81.78	0.43	83.04	0.41
	50	1.23	79.42	0.35	82.73	0.35	82.56	0.35	81.42	0.34	76.36	0.35
		0.91	94.16	0.35	95.07	0.35	94.22	0.35	93.90	0.35	93.52	0.35
		0.63	88.54	0.31	91.50	0.31	92.72	0.32	90.62	0.32	87.38	0.32
		0.42	91.12	0.26	93.60	0.27	94.14	0.28	92.28	0.27	91.80	0.27
		0.27	88.88	0.21	91.30	0.21	94.20	0.24	91.94	0.23	92.28	0.23
	100	1.23	67.10	0.25	68.17	0.25	70.50	0.25	69.44	0.25	64.82	0.25
		0.91	93.90	0.25	94.40	0.25	94.78	0.25	94.64	0.25	94.76	0.25
		0.63	85.48	0.22	88.07	0.22	90.94	0.23	89.20	0.23	86.36	0.23
		0.42	91.20	0.19	93.13	0.19	94.56	0.20	93.74	0.20	93.08	0.20
		0.27	90.56	0.15	91.50	0.15	94.26	0.17	93.58	0.17	93.94	0.17
	250	1.23	33.94	0.16	35.73	0.16	37.67	0.16	37.14	0.16	33.96	0.16
		0.91	94.30	0.16	94.60	0.16	93.86	0.16	93.84	0.16	94.38	0.13
		0.63	77.50	0.14	81.10	0.14	83.18	0.15	81.04	0.15	77.88	0.15
		0.42	90.98	0.12	92.83	0.12	93.64	0.13	92.78	0.13	91.92	0.16
		0.27	90.94	0.09	91.70	0.09	94.58	0.11	93.72	0.11	94.04	0.16

GCI - generalized confidence interval; BCa - bias corrected and accelerated; BP - basic percentile;
AN - asymptotic normal; Cov - coverage; Len - length

Table B.3: Coverage probability and length for 95% parametric CIs around θ_1^* under equal costs and three classes with a normally distributed feature.

			Delta		GCI		BCa		BP		AN	
	n_j	BC_3	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len
Normal	10	1.23	91.40	4.95	97.20	2.42	87.6	23.4	93.54	43.0	93.08	1.87
$\sigma_3 = 1$		0.91	92.50	1.04	97.10	1.40	91.24	2.44	92.36	5.52	93.04	1.06
		0.63	92.04	0.91	97.30	1.14	92.74	0.97	92.56	1.07	92.24	0.95
		0.42	92.20	0.94	96.70	1.10	93.32	1.02	92.74	1.00	92.40	1.00
		0.27	92.26	1.03	96.07	1.14	93.32	1.13	92.68	1.09	92.48	1.11
	50	1.23	93.80	0.58	95.73	0.63	92.32	0.58	92.68	1.29	93.88	0.58
		0.91	94.56	0.43	95.50	0.46	93.48	0.43	93.84	0.43	94.24	0.43
		0.63	94.10	0.39	94.97	0.41	94.14	0.40	94.36	0.40	94.56	0.40
		0.42	94.20	0.41	94.83	0.42	94.60	0.42	94.42	0.42	94.52	0.42
		0.27	94.78	0.46	95.03	0.46	94.58	0.46	94.40	0.46	94.18	0.46
	100	1.23	94.58	0.41	95.10	0.42	93.56	0.41	93.64	0.40	94.58	0.41
		0.91	94.76	0.30	95.37	0.31	94.14	0.30	94.38	0.30	94.60	0.30
		0.63	95.08	0.28	95.37	0.28	94.28	0.28	94.46	0.28	94.32	0.28
		0.42	95.12	0.29	95.30	0.29	94.40	0.29	94.24	0.29	94.16	0.29
		0.27	94.92	0.32	95.27	0.32	94.40	0.32	94.28	0.32	94.16	0.32
	250	1.23	94.60	0.26	94.90	0.26	94.94	0.26	94.92	0.25	95.00	0.26
		0.91	95.00	0.19	95.13	0.19	94.94	0.19	94.98	0.19	94.94	0.19
		0.63	94.80	0.18	94.97	0.18	94.68	0.18	94.90	0.18	94.68	0.18
		0.42	95.20	0.18	95.27	0.18	94.84	0.19	94.82	0.18	94.80	0.19
		0.27	95.16	0.20	95.33	0.20	94.70	0.20	94.52	0.20	94.66	0.20
Normal	10	1.23	92.22	10.0	97.20	2.42	87.58	23.4	93.54	43.0	93.08	1.87
$\sigma_3 = 2$		0.91	93.28	1.23	97.10	1.40	91.24	2.44	92.36	5.52	93.04	1.06
		0.63	92.86	0.91	97.30	1.14	92.74	0.97	92.56	1.07	92.24	0.95
		0.42	92.32	0.94	96.70	1.10	93.32	1.02	92.74	1.00	92.40	1.00
		0.27	92.22	1.03	96.07	1.14	93.32	1.13	92.68	1.09	92.48	1.11
	50	1.23	93.84	0.58	95.73	0.63	92.32	0.58	92.68	1.29	93.88	0.58
		0.91	94.54	0.43	95.50	0.46	93.48	0.43	93.84	0.43	94.24	0.43
		0.63	94.10	0.39	94.97	0.41	94.14	0.40	94.36	0.40	94.56	0.40
		0.42	94.18	0.41	94.83	0.42	94.60	0.42	94.42	0.42	94.52	0.42
		0.27	94.24	0.46	95.03	0.46	94.58	0.46	94.40	0.46	94.18	0.46
	100	1.23	94.66	0.41	95.10	0.42	93.56	0.41	93.64	0.40	94.58	0.41
		0.91	95.04	0.30	95.37	0.31	94.14	0.30	94.38	0.30	94.60	0.30
		0.63	95.08	0.28	95.37	0.28	94.28	0.28	94.46	0.28	94.32	0.28
		0.42	95.14	0.29	95.30	0.29	94.40	0.29	94.24	0.29	94.16	0.29
		0.27	94.92	0.32	95.27	0.32	94.40	0.32	94.28	0.32	94.16	0.32
	250	1.23	94.58	0.26	94.90	0.26	94.94	0.26	94.92	0.25	95.00	0.26

Continued on next page

Table B.3 – continued from previous page

			Delta		GCI		BCa		BP		AN	
	n_j	BC_3	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len
		0.91	94.64	0.19	95.13	0.19	94.94	0.19	94.98	0.19	94.94	0.19
		0.63	94.80	0.18	94.97	0.18	94.68	0.18	94.90	0.18	94.68	0.18
		0.42	95.20	0.18	95.27	0.18	94.84	0.19	94.82	0.18	94.80	0.19
		0.27	95.18	0.20	95.33	0.20	94.70	0.20	94.52	0.20	94.66	0.20
Normal	10	1.23	92.14	11.3	97.20	2.42	87.58	23.4	93.54	43.0	93.08	1.87
$\sigma_3 = 4$		0.91	93.28	1.24	97.10	1.40	91.24	2.44	92.36	5.52	93.04	1.06
		0.63	92.86	0.91	97.30	1.14	92.74	0.97	92.56	1.07	92.24	0.95
		0.42	92.36	0.94	96.70	1.10	93.32	1.02	92.74	1.00	92.40	1.00
		0.27	92.22	1.03	96.07	1.14	93.32	1.13	92.68	1.09	92.48	1.11
	50	1.23	93.84	0.58	95.73	0.63	92.32	0.58	92.68	1.29	93.88	0.58
		0.91	94.54	0.43	95.50	0.46	93.48	0.43	93.84	0.43	94.24	0.43
		0.63	94.10	0.39	94.97	0.41	94.14	0.40	94.36	0.40	94.56	0.40
		0.42	94.20	0.41	94.83	0.42	94.60	0.42	94.42	0.42	94.52	0.42
		0.27	94.24	0.46	95.03	0.46	94.58	0.46	94.40	0.46	94.18	0.46
	100	1.23	94.68	0.41	95.10	0.42	93.56	0.41	93.64	0.40	94.58	0.41
		0.91	95.04	0.30	95.37	0.31	94.14	0.30	94.38	0.30	94.60	0.30
		0.63	95.08	0.28	95.37	0.28	94.28	0.28	94.46	0.28	94.32	0.28
		0.42	95.12	0.29	95.30	0.29	94.40	0.29	94.24	0.29	94.16	0.29
		0.27	94.92	0.32	95.27	0.32	94.40	0.32	94.28	0.32	94.16	0.32
	250	1.23	94.60	0.26	94.90	0.26	94.94	0.26	94.92	0.25	95.00	0.26
		0.91	94.64	0.19	95.13	0.19	94.94	0.19	94.98	0.19	94.94	0.19
		0.63	94.80	0.18	94.97	0.18	94.68	0.18	94.90	0.18	94.68	0.18
		0.42	95.20	0.18	95.27	0.18	94.84	0.19	94.82	0.18	94.80	0.19
		0.27	95.18	0.20	95.33	0.20	94.70	0.20	94.52	0.20	94.66	0.20

GCI - generalized confidence interval; BCa - bias corrected and accelerated; BP - basic percentile;
AN - asymptotic normal; Cov - coverage; Len - length

Table B.4: Coverage probability and length for 95% parametric CIs around θ_2^* under equal costs and three classes with a normally distributed feature.

			Delta		GCI		BCa		BP		AN	
	n_j	BC_3	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len
Normal	10	1.23	91.52	1.85	97.00	2.46	87.92	43.7	93.38	20.7	92.56	1.90
$\sigma_3 = 1$		0.91	92.42	1.09	96.73	1.40	91.10	1.28	92.24	3.16	92.56	1.06
		0.63	92.32	0.91	96.33	1.14	92.76	0.97	92.74	1.11	92.42	0.95
		0.42	92.36	0.94	95.80	1.09	93.56	1.02	92.96	0.99	92.48	1.01
		0.27	92.48	1.03	95.53	1.13	94.00	1.13	92.92	1.09	92.34	1.11
	50	1.23	93.96	0.58	95.97	0.64	92.34	0.58	92.74	0.62	94.28	0.58
		0.91	94.60	0.43	95.57	0.46	93.56	0.43	93.92	0.43	94.58	0.43
		0.63	94.74	0.39	95.23	0.41	94.72	0.40	94.78	0.40	94.68	0.40
		0.42	94.58	0.41	94.83	0.42	94.72	0.42	94.66	0.42	94.42	0.42
		0.27	94.16	0.46	94.47	0.46	94.62	0.46	94.36	0.46	94.20	0.46
	100	1.23	94.46	0.41	95.13	0.42	93.22	0.41	93.48	0.40	94.26	0.40
		0.91	94.80	0.30	95.70	0.31	94.10	0.30	94.40	0.30	94.58	0.30
		0.63	94.94	0.28	95.03	0.28	94.42	0.28	94.54	0.28	94.56	0.28
		0.42	95.00	0.29	94.93	0.29	94.68	0.29	94.84	0.29	94.76	0.29
		0.27	95.18	0.32	95.00	0.32	95.14	0.32	95.10	0.32	95.08	0.32
	250	1.23	95.40	0.26	95.20	0.26	94.34	0.26	94.66	0.26	94.84	0.26
		0.91	94.90	0.19	95.27	0.19	94.70	0.19	94.86	0.19	94.92	0.19
		0.63	95.20	0.18	95.20	0.18	95.08	0.18	95.20	0.18	95.38	0.18
		0.42	95.12	0.18	95.40	0.18	95.20	0.19	95.20	0.18	95.22	0.18
		0.27	94.88	0.20	95.33	0.20	94.92	0.20	94.96	0.20	94.96	0.20
Normal	10	1.23	91.54	17.0	96.43	2.88	89.24	179	92.20	80.3	88.28	2.26
$\sigma_3 = 2$		0.91	92.26	9.82	96.67	1.76	90.24	1.85	91.36	5.81	89.02	1.37
		0.63	92.64	3.53	96.27	1.49	91.74	1.31	91.58	1.89	90.34	1.27
		0.42	92.24	2.18	95.80	1.47	92.58	1.38	92.20	1.34	91.14	1.35
		0.27	92.14	1.41	95.23	1.55	93.42	1.54	92.28	1.48	92.06	1.50
	50	1.23	94.22	0.66	95.27	0.69	92.84	0.67	93.38	0.67	92.78	0.66
		0.91	94.28	0.57	95.20	0.59	93.16	0.58	93.04	0.57	92.98	0.57
		0.63	94.18	0.55	94.53	0.56	93.86	0.55	93.82	0.54	93.52	0.55
		0.42	94.48	0.57	94.53	0.58	94.14	0.57	93.92	0.57	93.68	0.57
		0.27	94.62	0.62	94.60	0.63	93.82	0.63	93.86	0.62	93.58	0.62
	100	1.23	95.06	0.47	94.97	0.47	93.76	0.47	93.60	0.46	93.24	0.46
		0.91	95.16	0.41	95.07	0.41	93.90	0.41	94.00	0.40	93.44	0.40
		0.63	95.20	0.39	94.90	0.39	94.58	0.39	94.24	0.39	94.14	0.39
		0.42	95.26	0.40	94.73	0.41	94.80	0.40	94.64	0.40	94.72	0.40
		0.27	95.34	0.44	94.93	0.44	95.26	0.44	95.10	0.44	95.14	0.44
	250	1.23	94.82	0.29	95.50	0.30	94.88	0.30	94.88	0.29	94.98	0.29

Continued on next page

Table B.4 – continued from previous page

			Delta		GCI		BCa		BP		AN	
	n_j	BC_3	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len
		0.91	94.86	0.26	95.33	0.26	95.24	0.26	95.16	0.26	95.20	0.26
		0.63	95.18	0.24	94.83	0.25	95.22	0.25	95.14	0.24	95.20	0.25
		0.42	95.24	0.25	95.07	0.25	95.50	0.25	95.28	0.25	95.08	0.25
		0.27	95.08	0.28	94.73	0.28	95.20	0.28	95.04	0.28	94.96	0.28
Normal	10	1.23	91.10	24.7	96.10	2.35	89.30	14.7	90.66	11.4	84.70	1.95
$\sigma_3 = 4$		0.91	91.46	6.41	95.97	1.92	89.00	2.14	89.66	2.58	84.64	1.55
		0.63	92.06	1.59	95.93	1.84	90.30	1.65	90.16	1.69	87.00	1.54
		0.42	92.14	1.65	96.10	1.89	91.66	1.75	91.04	1.67	88.94	1.66
		0.27	92.24	1.79	95.60	2.02	92.86	1.95	91.96	1.84	90.04	1.85
	50	1.23	94.32	0.75	95.23	0.78	92.96	0.76	92.90	0.74	92.14	0.74
		0.91	93.80	0.71	95.13	0.73	92.86	0.71	93.00	0.69	91.84	0.70
		0.63	94.18	0.71	94.97	0.72	93.12	0.70	93.46	0.69	92.42	0.69
		0.42	94.20	0.73	95.03	0.75	93.30	0.73	93.28	0.72	92.84	0.72
		0.27	94.34	0.78	95.07	0.80	93.48	0.79	93.50	0.77	93.00	0.78
	100	1.23	94.60	0.53	94.60	0.54	94.08	0.54	93.38	0.53	92.56	0.53
		0.91	95.00	0.50	94.93	0.51	94.00	0.51	93.56	0.50	92.98	0.50
		0.63	95.28	0.50	95.00	0.50	94.18	0.50	93.92	0.49	94.00	0.50
		0.42	95.32	0.52	94.93	0.52	94.30	0.52	94.64	0.51	94.22	0.51
		0.27	95.32	0.55	94.87	0.56	94.86	0.56	94.94	0.55	94.78	0.55
	250	1.23	95.06	0.34	95.43	0.34	94.96	0.34	94.82	0.34	94.68	0.34
		0.91	94.82	0.32	95.83	0.32	94.94	0.32	94.88	0.32	94.66	0.32
		0.63	95.16	0.32	95.00	0.32	94.66	0.32	94.76	0.31	94.78	0.32
		0.42	95.22	0.33	94.90	0.33	94.82	0.33	94.80	0.32	94.56	0.33
		0.27	95.10	0.35	94.60	0.35	94.58	0.35	94.88	0.35	94.60	0.35

GCI - generalized confidence interval; BCa - bias corrected and accelerated; BP - basic percentile;
AN - asymptotic normal; Cov - coverage; Len - length

Table B.5: Coverage probability and length for 95% parametric CIs around θ_1^* under equal costs and three classes with a non-normally distributed feature.

			Delta		GCI		BCa		BP		AN	
	n_j	BC_3	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len
Gamma	10	1.23	54.30	1.67	65.33	2.76	55.14	1.77	53.56	1.80	83.40	1.78
		0.91	55.28	4.66	65.33	2.76	55.14	1.77	53.56	1.80	83.40	1.78
		0.63	55.44	1.76	64.50	2.84	55.26	1.81	53.84	1.84	83.58	1.81
		0.42	72.24	1.87	74.67	2.18	81.78	2.48	80.20	2.45	85.88	2.38
		0.27	72.18	1.87	74.67	2.18	81.78	2.48	80.20	2.45	85.88	2.38
	50	1.23	1.58	0.71	2.37	0.75	1.94	0.94	1.64	0.91	6.88	0.92
		0.91	1.58	0.71	2.37	0.75	1.94	0.94	1.64	0.91	6.88	0.92
		0.63	1.58	0.71	1.97	0.74	2.08	0.93	1.80	0.91	7.02	0.91
		0.42	37.30	0.87	33.10	0.89	56.52	1.43	62.30	1.37	74.84	1.37
		0.27	37.28	0.87	33.10	0.89	56.52	1.43	62.30	1.37	74.84	1.37
	100	1.23	0.00	0.49	0.07	0.50	0.28	0.67	0.06	0.64	0.48	0.64
		0.91	0.00	0.49	0.07	0.50	0.28	0.67	0.06	0.64	0.48	0.64
		0.63	0.08	0.49	0.03	0.50	0.18	0.67	0.08	0.64	0.48	0.64
		0.42	15.88	0.62	14.53	0.62	30.18	1.06	37.08	1.02	45.58	1.02
		0.27	15.92	0.62	14.53	0.62	30.18	1.06	37.08	1.02	45.58	1.02
	250	1.23	0.00	0.31	0.00	0.31	0.00	0.42	0.00	0.41	0.00	0.41
		0.91	0.00	0.31	0.00	0.31	0.00	0.42	0.00	0.41	0.00	0.41
		0.63	0.00	0.31	0.00	0.31	0.00	0.42	0.00	0.41	0.00	0.41
		0.42	1.00	0.39	0.73	0.39	4.16	0.70	6.38	0.68	8.10	0.68
		0.27	1.10	0.39	0.73	0.39	4.16	0.70	6.38	0.68	8.10	0.68
Gamma w/ Box-Cox	10	1.23	89.02	2.11	95.10	63.7	88.90	1.62	87.68	1.65	93.50	1.64
		0.91	84.24	3.13	94.77	122	84.64	1.52	82.14	1.63	91.56	1.62
		0.63	78.60	242	91.47	567	78.18	1.43	74.60	1.64	89.60	1.63
		0.42	91.12	2.19	95.93	2.59	91.64	2.39	88.96	2.38	91.40	2.38
		0.27	89.10	2.17	94.57	2.55	89.84	2.35	86.00	2.42	90.74	2.41
	50	1.23	94.06	0.73	94.93	0.79	93.94	0.75	92.88	0.74	94.18	0.74
		0.91	73.76	0.68	75.60	0.73	71.58	0.69	65.70	0.69	75.70	0.70
		0.63	37.12	0.60	40.90	0.66	33.72	0.63	28.64	0.65	43.46	0.65
		0.42	86.80	0.94	87.67	0.97	88.82	1.00	85.58	1.00	89.24	1.00
		0.27	73.68	0.92	74.30	0.95	74.98	0.99	70.62	0.99	79.46	1.00
	100	1.23	94.12	0.51	94.97	0.53	94.26	0.52	93.74	0.51	94.38	0.52
		0.91	53.18	0.47	57.20	0.49	52.32	0.48	46.16	0.47	55.18	0.48
		0.63	10.20	0.43	10.13	0.44	10.04	0.43	7.44	0.44	13.26	0.44
		0.42	79.14	0.66	79.27	0.67	81.88	0.71	79.00	0.71	82.44	0.71
		0.27	53.36	0.65	51.47	0.66	56.88	0.70	53.04	0.70	60.14	0.70
	250	1.23	93.32	0.32	93.67	0.33	93.88	0.32	92.80	0.32	93.76	0.32

Continued on next page

Table B.5 – continued from previous page

			Delta		GCI		BCa		BP		AN	
	n_j	BC_3	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len
		0.91	15.38	0.30	16.27	0.30	15.40	0.30	12.74	0.29	16.22	0.30
		0.63	0.10	0.27	0.17	0.27	0.20	0.26	0.10	0.26	0.22	0.27
		0.42	57.02	0.42	57.07	0.42	61.22	0.45	58.58	0.45	61.94	0.45
		0.27	15.76	0.41	15.10	0.41	18.26	0.44	16.74	0.44	19.82	0.44
Normal	10	1.23	87.10	1.28	95.37	6.03	91.80	1.46	89.46	1.48	87.48	1.42
Mixture		0.91	87.02	1.28	95.97	1.85	91.80	1.46	89.46	1.48	87.48	1.42
		0.63	87.02	1.28	94.13	1.47	91.80	1.46	89.46	1.48	87.48	1.42
		0.42	90.96	1.32	94.13	1.47	89.00	1.99	90.50	5.14	88.74	1.35
		0.27	89.56	90.2	94.13	1.47	85.88	159	94.10	381	90.50	5.27
	50	1.23	76.34	0.56	90.40	0.97	78.20	0.56	76.34	0.56	74.38	0.56
		0.91	76.34	0.56	93.33	0.59	78.20	0.56	76.34	0.56	74.36	0.56
		0.63	76.34	0.56	80.93	0.57	78.20	0.56	76.34	0.56	74.36	0.56
		0.42	91.10	0.57	80.93	0.57	90.80	0.57	90.04	0.55	89.08	0.56
		0.27	92.30	0.84	80.93	0.57	90.64	1.06	93.20	4.68	93.42	0.99
	100	1.23	57.86	0.39	87.13	0.59	60.10	0.39	58.00	0.39	55.64	0.39
		0.91	57.86	0.39	90.67	0.41	60.10	0.39	58.00	0.39	55.64	0.39
		0.63	57.86	0.39	61.93	0.40	60.10	0.39	58.00	0.39	55.64	0.39
		0.42	89.64	0.40	61.93	0.40	89.34	0.40	87.90	0.40	87.26	0.40
		0.27	88.00	0.57	61.93	0.40	87.80	0.61	88.90	0.63	90.76	0.61
	250	1.23	20.26	0.25	75.53	0.36	21.92	0.25	20.32	0.24	19.66	0.25
		0.91	20.26	0.25	84.27	0.26	21.92	0.25	20.32	0.24	19.66	0.25
		0.63	20.26	0.25	21.77	0.25	21.92	0.25	20.32	0.24	19.66	0.25
		0.42	83.08	0.25	21.77	0.25	84.82	0.25	82.66	0.25	82.12	0.25
		0.27	77.98	0.35	21.77	0.25	78.30	0.38	78.96	0.37	82.56	0.38

GCI - generalized confidence interval; BCa - bias corrected and accelerated; BP - basic percentile;
AN - asymptotic normal; Cov - coverage; Len - length

Table B.6: Coverage probability and length for 95% parametric CIs around θ_2^* under equal costs and three classes with a non-normally distributed feature.

			Delta		GCI		BCa		BP		AN	
	n_j	BC_3	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len
Gamma	10	1.23	66.20	3.67	78.17	6.95	58.90	3.03	58.22	3.17	85.56	3.17
		0.91	71.68	2.92	75.93	3.44	76.18	3.62	75.60	3.58	87.80	3.52
		0.63	70.54	4.49	82.67	5.10	79.32	5.82	76.02	5.60	73.00	5.48
		0.42	79.64	7.73	82.70	8.92	84.58	9.66	83.42	9.47	88.34	9.37
		0.27	74.26	11.7	83.27	13.3	81.34	15.0	78.50	14.4	75.52	14.1
	50	1.23	11.74	1.47	15.10	1.60	14.22	1.70	11.50	1.71	26.92	1.73
		0.91	16.02	1.33	15.03	1.36	24.64	1.92	28.68	1.86	41.76	1.86
		0.63	61.82	2.00	65.10	2.04	77.92	2.91	74.80	2.83	71.58	2.81
		0.42	46.80	3.52	43.23	3.59	59.20	4.82	62.96	4.72	74.24	4.73
		0.27	66.76	5.20	72.20	5.32	82.02	7.40	78.60	7.20	75.80	7.17
	100	1.23	1.64	1.02	2.13	1.06	4.24	1.18	2.30	1.18	6.82	1.19
		0.91	2.00	0.95	2.07	0.96	4.94	1.41	6.54	1.37	9.88	1.37
		0.63	49.78	1.41	54.30	1.43	75.16	2.16	70.12	2.10	67.40	2.10
		0.42	21.84	2.49	20.50	2.52	33.64	3.52	38.20	3.46	47.04	3.46
		0.27	57.92	3.68	61.13	3.71	79.58	5.44	75.06	5.29	72.74	5.29
	250	1.23	0.00	0.63	0.00	0.64	0.08	0.73	0.02	0.72	0.10	0.72
		0.91	0.00	0.60	0.00	0.60	0.02	0.91	0.04	0.90	0.04	0.90
		0.63	27.62	0.90	30.00	0.90	57.86	1.41	51.14	1.37	49.56	1.38
		0.42	1.94	1.58	1.77	1.59	4.96	2.29	6.00	2.26	7.58	2.26
		0.27	39.44	2.33	42.33	2.34	67.82	3.55	61.84	3.48	59.84	3.49
Gamma w/ Box-Cox	10	1.23	89.58	3.65	92.60	999	84.64	2.85	85.46	3.03	92.92	3.02
		0.91	88.90	3.06	95.07	3.89	90.44	3.33	87.54	3.35	91.80	3.35
		0.63	90.02	6.92	93.30	7.14	90.84	7.41	92.56	7.63	87.66	7.63
		0.42	91.32	8.27	95.23	9.61	92.16	9.02	90.68	9.03	91.94	9.07
		0.27	90.92	17.3	94.77	18.0	92.32	18.6	92.32	18.7	90.56	18.8
	50	1.23	94.76	1.42	95.13	1.56	93.42	1.42	92.98	1.41	94.64	1.42
		0.91	78.50	1.31	79.50	1.36	81.14	1.42	78.84	1.42	81.94	1.42
		0.63	88.38	3.00	90.40	3.00	89.94	3.07	91.30	3.09	87.54	3.10
		0.42	90.76	3.63	91.93	3.70	90.72	3.79	89.84	3.78	91.94	3.80
		0.27	92.00	7.50	93.00	7.51	91.70	7.55	93.08	7.58	90.82	7.61
	100	1.23	94.06	1.00	95.03	1.05	93.20	0.99	92.56	0.99	93.98	0.99
		0.91	64.18	0.93	66.17	0.94	68.18	1.01	66.26	1.00	68.76	1.01
		0.63	84.88	2.11	86.93	2.11	87.02	2.17	88.42	2.18	84.46	2.18
		0.42	86.44	2.56	86.23	2.58	87.32	2.67	86.06	2.66	88.54	2.68
		0.27	89.60	5.28	90.43	5.29	90.26	5.37	91.26	5.37	89.04	5.39
	250	1.23	93.70	0.63	94.20	0.64	93.24	0.62	92.90	0.62	93.38	0.62

Continued on next page

Table B.6 – continued from previous page

			Delta		GCI		BCa		BP		AN	
	n_j	BC_3	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len
		0.91	31.34	0.58	34.57	0.59	35.56	0.63	34.16	0.63	35.74	0.63
		0.63	75.20	1.33	75.07	1.33	77.96	1.38	79.12	1.38	75.46	1.38
		0.42	74.76	1.62	74.57	1.62	76.52	1.69	75.28	1.68	77.96	1.69
		0.27	84.10	3.34	83.50	3.34	85.40	3.41	86.48	3.40	83.86	3.42
Normal	10	1.23	92.32	1.22	96.77	1.55	92.10	1.36	92.58	2.47	90.04	1.26
Mixture		0.91	92.22	1.22	96.77	1.55	92.10	1.36	92.58	2.47	90.04	1.26
		0.63	92.16	1.13	95.97	1.40	92.22	1.21	91.94	1.54	91.70	1.18
		0.42	92.52	1.23	94.70	1.37	92.30	1.35	92.68	1.30	91.94	1.32
		0.27	91.78	2.08	95.20	2.01	92.22	2.07	90.00	1.98	90.46	2.03
	50	1.23	95.68	0.52	95.07	0.54	94.52	0.52	94.66	0.51	94.12	0.51
		0.91	95.68	0.52	95.07	0.54	94.52	0.52	94.66	0.51	94.12	0.51
		0.63	92.16	0.49	90.97	0.51	90.32	0.49	90.58	0.49	92.18	0.49
		0.42	88.16	0.55	88.50	0.55	87.24	0.55	87.88	0.54	89.44	0.54
		0.27	91.52	0.86	91.73	0.87	90.26	0.82	88.88	0.81	88.44	0.82
	100	1.23	95.54	0.37	95.70	0.37	94.88	0.36	94.78	0.36	94.88	0.36
		0.91	95.52	0.37	95.70	0.37	94.88	0.36	94.78	0.36	94.88	0.36
		0.63	87.54	0.35	86.80	0.35	85.44	0.35	86.06	0.34	88.14	0.35
		0.42	82.02	0.38	79.90	0.39	78.58	0.38	79.72	0.38	81.76	0.38
		0.27	86.82	0.61	88.23	0.61	85.70	0.58	84.46	0.57	83.88	0.58
	250	1.23	95.40	0.23	95.73	0.23	94.96	0.23	94.80	0.23	95.00	0.23
		0.91	95.40	0.23	95.73	0.23	94.96	0.23	94.80	0.23	95.00	0.23
		0.63	74.44	0.22	72.87	0.22	72.24	0.22	72.94	0.22	75.40	0.22
		0.42	59.00	0.24	58.47	0.24	56.84	0.24	58.06	0.24	60.10	0.24
		0.27	74.10	0.38	76.23	0.38	72.88	0.37	71.42	0.36	70.72	0.37

GCI - generalized confidence interval; BCa - bias corrected and accelerated; BP - basic percentile;
AN - asymptotic normal; Cov - coverage; Len - length

Table B.7: Coverage probability and length for 95% parametric CIs around *BC* under unequal costs and and three classes with a normally distributed feature.

			Delta		GCI		BCa		BP		AN	
	n_j	BC_3	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len
$Cost_1$	10	0.45	91.10	0.31	96.73	0.31	91.78	0.30	89.84	0.30	87.54	0.30
		0.31	93.04	0.27	95.60	0.28	91.48	0.26	90.66	0.25	86.88	0.25
		0.21	92.18	0.23	96.27	0.24	90.90	0.22	88.66	0.21	86.26	0.20
		0.14	90.84	0.19	95.37	0.21	90.40	0.19	86.10	0.17	85.12	0.16
		0.09	89.94	0.16	94.87	0.18	89.94	0.16	83.90	0.13	85.26	0.12
	50	0.45	94.50	0.14	95.37	0.14	95.14	0.14	94.46	0.14	93.54	0.14
		0.31	94.68	0.12	94.97	0.12	94.42	0.12	94.72	0.12	93.72	0.12
		0.21	94.64	0.10	94.90	0.10	94.30	0.10	94.08	0.10	93.16	0.10
		0.14	94.08	0.09	94.80	0.09	94.14	0.09	93.34	0.08	93.24	0.08
		0.09	93.56	0.07	95.07	0.07	94.00	0.07	92.64	0.07	93.20	0.07
	100	0.45	95.16	0.10	95.07	0.10	94.60	0.10	94.20	0.10	94.06	0.10
		0.31	95.20	0.08	94.37	0.08	94.42	0.09	94.54	0.08	93.94	0.08
		0.21	95.00	0.07	94.80	0.07	94.38	0.07	94.12	0.07	93.94	0.07
		0.14	94.74	0.06	95.07	0.06	94.14	0.06	93.72	0.06	93.90	0.06
		0.09	94.70	0.05	95.10	0.05	94.14	0.05	93.80	0.05	93.60	0.05
	250	0.45	94.92	0.06	94.83	0.06	95.12	0.06	95.06	0.06	94.96	0.06
		0.31	94.74	0.05	94.43	0.05	94.80	0.05	94.64	0.05	94.84	0.05
		0.21	94.78	0.05	94.43	0.05	94.82	0.05	94.32	0.05	94.52	0.05
		0.14	94.78	0.04	95.27	0.04	94.54	0.04	94.28	0.04	94.60	0.04
		0.09	94.68	0.03	94.97	0.03	94.36	0.03	94.24	0.03	94.52	0.03
$Cost_2$	10	0.89	91.46	0.63	95.13	0.61	91.26	0.59	89.16	0.61	85.76	0.60
		0.66	91.90	0.60	96.63	0.60	92.30	0.58	88.90	0.56	85.24	0.56
		0.46	91.04	0.53	96.50	0.55	90.80	0.54	87.50	0.48	83.92	0.48
		0.31	90.02	0.45	96.40	0.48	90.46	0.47	85.32	0.39	83.54	0.38
		0.20	89.00	0.36	95.97	0.41	89.54	0.39	83.02	0.30	83.46	0.29
	50	0.89	94.18	0.29	94.77	0.29	94.02	0.29	93.74	0.29	93.24	0.29
		0.66	94.28	0.27	96.00	0.27	94.14	0.27	93.82	0.27	93.16	0.27
		0.46	94.18	0.24	95.57	0.24	94.02	0.24	93.20	0.24	92.94	0.24
		0.31	93.84	0.21	95.17	0.21	93.84	0.21	92.76	0.20	92.82	0.20
		0.20	93.28	0.17	94.70	0.17	93.58	0.17	92.30	0.16	92.76	0.16
	100	0.89	94.86	0.21	94.47	0.21	93.94	0.21	93.90	0.21	93.66	0.21
		0.66	94.78	0.19	95.10	0.19	94.20	0.19	93.98	0.19	93.92	0.19
		0.46	94.60	0.17	94.70	0.17	94.20	0.17	93.82	0.17	93.62	0.17
		0.31	94.26	0.15	95.00	0.15	94.54	0.15	93.66	0.14	93.58	0.14
		0.20	93.88	0.12	95.00	0.12	94.32	0.12	93.54	0.11	93.74	0.11
	250	0.89	94.62	0.13	94.93	0.13	95.02	0.13	94.92	0.13	94.56	0.13

Continued on next page

Table B.7 – continued from previous page

		Delta		GCI		BCa		BP		AN	
n_j	BC_3	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len
	0.66	94.68	0.12	95.00	0.12	94.80	0.12	94.74	0.12	94.36	0.12
	0.46	94.58	0.11	95.13	0.11	94.68	0.11	94.56	0.11	94.36	0.11
	0.31	94.74	0.09	94.77	0.09	94.56	0.09	94.54	0.09	94.50	0.09
	0.20	94.70	0.07	95.00	0.07	94.70	0.08	94.48	0.07	94.56	0.07

GCI - generalized confidence interval; BCa - bias corrected and accelerated; BP - basic percentile; AN - asymptotic normal; Cov - coverage; Len - length

Table B.8: Coverage probability and length for 95% parametric CIs around θ_1^* under unequal costs and three classes with a normally distributed feature.

			Delta		GCI		BCa		BP		AN	
	n_j	BC_3	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len
$Cost_1$	10	0.45	91.28	52.0	97.00	4.33	90.08	6.75	94.74	4.78	92.68	2.38
		0.31	93.42	0.93	96.73	1.44	92.24	1.23	93.82	1.22	93.84	0.99
		0.21	93.04	0.89	97.13	1.10	93.00	0.95	93.62	0.95	92.52	0.93
		0.14	92.30	0.93	95.43	1.07	93.40	1.01	92.92	0.99	92.48	1.00
		0.09	92.26	1.02	96.07	1.13	93.34	1.12	92.80	1.09	92.52	1.10
	50	0.45	94.04	0.49	95.73	0.54	93.36	0.52	93.30	0.53	93.22	0.50
		0.31	95.02	0.40	95.50	0.41	93.80	0.40	93.90	0.39	94.22	0.40
		0.21	94.38	0.39	94.90	0.40	94.14	0.39	94.24	0.39	94.44	0.39
		0.14	94.10	0.41	94.57	0.42	94.50	0.42	94.38	0.41	94.60	0.42
		0.09	94.22	0.46	95.23	0.46	94.62	0.46	94.38	0.46	94.18	0.46
	100	0.45	94.80	0.35	94.90	0.36	94.16	0.35	94.16	0.34	94.42	0.35
		0.31	94.98	0.28	95.10	0.29	94.82	0.28	94.82	0.28	94.96	0.28
		0.21	95.02	0.27	95.20	0.28	94.28	0.27	94.36	0.27	94.32	0.27
		0.14	95.14	0.29	95.57	0.29	94.40	0.29	94.24	0.29	94.16	0.29
		0.09	94.92	0.32	94.87	0.32	94.38	0.32	94.28	0.32	94.14	0.32
	250	0.45	94.72	0.22	94.47	0.22	94.76	0.22	94.68	0.22	94.84	0.22
		0.31	94.64	0.18	95.83	0.18	94.78	0.18	94.96	0.18	94.86	0.18
		0.21	94.62	0.17	95.30	0.17	94.72	0.17	94.78	0.17	94.74	0.17
		0.14	95.18	0.18	94.43	0.18	94.88	0.18	94.88	0.18	94.78	0.18
		0.09	95.14	0.20	95.17	0.20	94.70	0.20	94.54	0.20	94.66	0.20
$Cost_2$	10	0.89	87.90	446	96.43	8.79	88.80	11.5	87.44	8.16	89.38	5.23
		0.66	92.22	49.5	96.97	5.26	89.94	4.94	93.26	3.56	89.48	2.89
		0.46	93.56	7.47	96.37	2.96	90.94	2.52	93.96	1.94	90.28	1.64
		0.31	93.04	3.13	97.10	1.81	92.32	1.46	93.74	1.30	90.84	1.16
		0.20	92.26	1.09	96.23	1.36	92.84	1.22	92.84	1.18	91.56	1.14
	50	0.89	93.78	15.9	95.60	2.63	94.18	2.97	94.70	2.69	93.30	2.27
		0.66	94.02	0.52	95.23	0.58	93.36	0.61	94.40	0.60	93.00	0.55
		0.46	93.96	0.45	94.73	0.47	93.24	0.46	93.42	0.45	93.34	0.45
		0.31	94.34	0.45	94.83	0.46	93.42	0.45	93.46	0.45	93.36	0.45
		0.20	94.12	0.48	95.07	0.49	93.60	0.48	93.42	0.48	93.56	0.48
	100	0.89	95.08	0.59	94.83	1.10	94.60	1.23	96.02	1.28	94.16	1.03
		0.66	94.88	0.37	95.03	0.38	93.84	0.37	94.74	0.37	94.18	0.37
		0.46	94.78	0.32	94.60	0.33	94.26	0.32	94.66	0.32	94.54	0.32
		0.31	95.04	0.32	95.40	0.32	94.22	0.32	94.42	0.32	94.34	0.32
		0.20	95.18	0.34	95.53	0.34	94.00	0.34	94.06	0.34	94.04	0.34
	250	0.89	94.98	0.34	94.93	0.36	95.22	0.37	96.24	0.38	95.14	0.36

Continued on next page

Table B.8 – continued from previous page

		Delta		GCI		BCa		BP		AN	
n_j	BC_3	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len
	0.66	94.86	0.23	94.43	0.23	94.84	0.23	95.38	0.23	95.08	0.23
	0.46	95.04	0.20	94.90	0.20	94.64	0.20	94.50	0.20	94.34	0.20
	0.31	95.38	0.20	94.93	0.20	94.40	0.20	94.42	0.20	94.32	0.20
	0.20	95.54	0.22	94.70	0.22	94.36	0.22	94.32	0.21	94.34	0.22

GCI - generalized confidence interval; BCa - bias corrected and accelerated; BP - basic percentile; AN - asymptotic normal; Cov - coverage; Len - length

Table B.9: Coverage probability and length for 95% parametric CIs around θ_2^* under unequal costs and three classes with a normally distributed feature.

			Delta		GCI		BCa		BP		AN	
	n_j	BC_3	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len
$Cost_1$	10	0.45	91.54	28.4	97.23	4.56	89.86	7.97	94.54	5.61	91.88	2.55
		0.31	93.44	0.93	97.23	1.50	91.82	1.35	94.02	1.35	93.02	1.03
		0.21	92.86	0.89	97.17	1.10	93.18	0.97	93.76	0.97	92.96	0.94
		0.14	92.70	0.93	95.53	1.07	93.74	1.02	93.22	1.00	92.60	1.00
		0.09	92.66	1.03	96.40	1.13	93.84	1.12	93.10	1.10	92.42	1.11
	50	0.45	94.30	0.49	95.73	0.53	93.22	0.53	93.56	0.53	93.32	0.50
		0.31	94.62	0.40	95.50	0.42	93.96	0.40	94.02	0.39	94.42	0.40
		0.21	94.66	0.39	94.90	0.40	94.46	0.39	94.68	0.39	94.58	0.39
		0.14	94.60	0.41	94.57	0.42	94.76	0.42	94.58	0.41	94.50	0.42
		0.09	94.52	0.46	95.23	0.46	94.64	0.46	94.36	0.46	94.20	0.46
	100	0.45	94.88	0.35	94.90	0.36	93.54	0.35	93.40	0.34	93.82	0.35
		0.31	95.08	0.28	95.10	0.29	94.40	0.28	94.54	0.28	94.62	0.28
		0.21	95.00	0.27	95.20	0.28	94.48	0.27	94.72	0.27	94.74	0.27
		0.14	94.98	0.29	95.57	0.29	94.64	0.29	94.80	0.29	94.78	0.29
		0.09	95.20	0.32	94.87	0.32	95.12	0.32	95.08	0.32	95.12	0.32
	250	0.45	94.86	0.22	94.47	0.22	94.44	0.22	94.56	0.22	94.76	0.22
		0.31	95.56	0.18	95.83	0.18	94.98	0.18	94.94	0.18	95.12	0.18
		0.21	95.42	0.17	95.30	0.17	94.90	0.17	95.04	0.17	95.24	0.17
		0.14	95.18	0.18	94.43	0.18	95.12	0.18	95.14	0.18	95.28	0.18
		0.09	94.84	0.20	95.17	0.20	94.92	0.20	94.96	0.20	94.96	0.20
$Cost_2$	10	0.89	92.40	21.6	98.03	2.95	87.92	3.64	95.64	3.43	93.98	2.17
		0.66	93.22	5.80	97.43	1.61	91.66	1.32	94.66	1.51	93.64	1.28
		0.46	92.70	1.71	96.57	1.20	93.28	1.07	94.34	1.16	93.18	1.03
		0.31	92.64	0.94	96.00	1.10	93.90	1.06	93.94	1.10	92.74	1.04
		0.27	92.60	1.03	95.67	1.14	94.04	1.14	93.42	1.15	92.40	1.12
	50	0.89	93.98	0.58	95.57	0.63	92.34	0.58	92.78	0.58	94.28	0.58
		0.66	94.68	0.43	96.57	0.46	93.56	0.43	93.92	0.43	94.58	0.43
		0.46	94.76	0.39	94.93	0.41	94.72	0.40	94.78	0.40	94.68	0.40
		0.31	94.58	0.41	94.73	0.42	94.72	0.42	94.66	0.42	94.42	0.42
		0.27	94.50	0.46	95.07	0.46	94.62	0.46	94.36	0.46	94.20	0.46
	100	0.89	95.10	0.41	95.40	0.42	93.22	0.41	93.48	0.40	94.26	0.40
		0.66	94.84	0.30	95.20	0.31	94.10	0.30	94.40	0.30	94.58	0.30
		0.46	94.94	0.28	94.63	0.28	94.42	0.28	94.54	0.28	94.56	0.28
		0.31	95.00	0.29	95.53	0.29	94.68	0.29	94.84	0.29	94.76	0.29
		0.27	95.18	0.32	94.50	0.32	95.14	0.32	95.10	0.32	95.08	0.32
	250	0.89	95.40	0.26	95.03	0.26	94.34	0.26	94.66	0.26	94.84	0.26

Continued on next page

Table B.9 – continued from previous page

		Delta		GCI		BCa		BP		AN	
n_j	BC_3	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len
	0.66	95.52	0.19	94.87	0.19	94.70	0.19	94.86	0.19	94.92	0.19
	0.46	95.20	0.18	94.90	0.18	95.08	0.18	95.20	0.18	95.38	0.18
	0.31	95.12	0.18	95.47	0.18	95.20	0.19	95.20	0.18	95.22	0.18
	0.27	94.88	0.20	94.93	0.20	94.92	0.20	94.96	0.20	94.96	0.20

GCI - generalized confidence interval; BCa - bias corrected and accelerated; BP - basic percentile; AN - asymptotic normal; Cov - coverage; Len - length

B.2 Additional CI performance results from Chapter 4

Table B.10: Simulation coverage probability and length for the nonparametric bootstrapped 95% CI around BC_2 for two classes with a normally distributed feature when all $c_{i|j}p_j = 1$, for $i \neq j$.

n_1	n_2	BC_2	Coverage	Length
5	5	0.6	90.30	0.47
		0.4	63.60	0.35
		0.2	0.00	0.17
		0.1	0.00	0.06
6	9	0.6	91.47	0.45
		0.4	84.90	0.35
		0.2	0.00	0.18
		0.1	0.00	0.07
10	10	0.6	92.93	0.48
		0.4	94.80	0.39
		0.2	7.40	0.22
		0.1	0.00	0.10
12	18	0.6	90.27	0.42
		0.4	93.63	0.35
		0.2	73.47	0.22
		0.1	0.00	0.11
20	20	0.6	91.33	0.42
		0.4	94.83	0.35
		0.2	90.5	0.24
		0.1	0.00	0.13
22	28	0.6	87.10	0.35
		0.4	91.97	0.30
		0.2	91.60	0.20
		0.1	1.63	0.12
30	30	0.6	92.47	0.37
		0.4	95.43	0.32
		0.2	95.37	0.22
		0.1	69.70	0.14

Appendix C: R Code

C.1 R Code

C.1.1 Delta Method 95% CIs .

```
#IMPUTS TO CHANGE##
p1<-#SET Prevelance Class 1
p2<-#SET Prevelance Class 2
p3<-#SET Prevelance Class 3
w21<-#SET COST 2|1
w31<-#SET COST 3|1
w12<-#SET COST 1|2
w32<-#SET COST 3|2
w13<-#SET COST 1|3
w23<-#SET COST 2|3
start<-c(-.1,0)
L<-c(-1000,-1000)
U<-c(1000,1000)
nx<-#SIZE Class 1
ny<-#SIZE Class 2
nz<-#SIZE Class 3
X<-#Vector of Values for Class 1
Y<-#Vector of Values for Class 2
Z<-#Vector of Values for Class 3
gmu1<-mean(X)
gmu2<-mean(Y)
gmu3<-mean(Z)
gsig1<-sd(X)
gsig2<-sd(Y)
gsig3<-sd(Z)
f<-function(par){(pnorm(par[2],gmu1,gsig1)-pnorm(par[1],gmu1,gsig1))*(p1*w21)+
(1-pnorm(par[2],gmu1,gsig1))*(p1*w31)+
(pnorm(par[1],gmu2,gsig2))*(p2*w12)+
(1-pnorm(par[2],gmu2,gsig2))*(p2*w32)+
(pnorm(par[1],gmu3,gsig3))*(p3*w13)+
(pnorm(par[2],gmu3,gsig3)-pnorm(par[1],gmu3,gsig3))*(p3*w23)}
x<-nlminb(start, f, lower = L, upper = U)
c1<-x$par[1]
c2<-x$par[2]
EBC<-x$objective
##ESTIMATE PARTIALS FOR THETA
g<-function(par){(pnorm(par[2],mux,sigx)-pnorm(par[1],mux,sigx))*(p1*w21)+
(1-pnorm(par[2],mux,sigx))*(p1*w31)+
(pnorm(par[1],muy,sigy))*(p2*w12)+
(1-pnorm(par[2],muy,sigy))*(p2*w32)+
(pnorm(par[1],muz,sigz))*(p3*w13)+
(pnorm(par[2],muz,sigz)-pnorm(par[1],muz,sigz))*(p3*w23)}
#Partial for Theta 1 & 2 wrt Mean 1
#start with +eppsilon
mux<-gmu1+.0001
muy<-gmu2
muz<-gmu3
sigx<-gsig1
sigy<-gsig2
sigz<-gsig3
x<-nlminb(start, g, lower = L, upper = U)
o1p<-x$par[1]
o2p<-x$par[2]
#now  - eppsilon
```

```
mux<-gmu1-.0001
muy<-gmu2
muz<-gmu3
sigx<-gsig1
sigy<-gsig2
sigz<-gsig3
x<-nlminb(start, g, lower = L, upper = U)
o1m<-x$par[1]
o2m<-x$par[2]
#Calc Partial
dc1m1<-(o1p-o1m)/.0002
dc2m1<-(o2p-o2m)/.0002
#Partial for Theta 1 & 2 wrt Mean 2
#start with +eppsilon
mux<-gmu1
muy<-gmu2+.0001
muz<-gmu3
sigx<-gsig1
sigy<-gsig2
sigz<-gsig3
x<-nlminb(start, g, lower = L, upper = U)
o1p<-x$par[1]
o2p<-x$par[2]
#now - eppsilon
mux<-gmu1
muy<-gmu2-.0001
muz<-gmu3
sigx<-gsig1
sigy<-gsig2
sigz<-gsig3
x<-nlminb(start, g, lower = L, upper = U)
o1m<-x$par[1]
o2m<-x$par[2]
#Calc Partial
dc1m2<-(o1p-o1m)/.0002
dc2m2<-(o2p-o2m)/.0002
#Partial for Theta 1 & 2 wrt Mean 3
#start with +eppsilon
mux<-gmu1
muy<-gmu2
muz<-gmu3+.0001
sigx<-gsig1
sigy<-gsig2
sigz<-gsig3
x<-nlminb(start, g, lower = L, upper = U)
o1p<-x$par[1]
o2p<-x$par[2]
#now - eppsilon
mux<-gmu1
muy<-gmu2
muz<-gmu3-.0001
sigx<-gsig1
sigy<-gsig2
sigz<-gsig3
x<-nlminb(start, g, lower = L, upper = U)
o1m<-x$par[1]
o2m<-x$par[2]
```

```
#Calc Partial
dc1m3<-(o1p-o1m)/.0002
dc2m3<-(o2p-o2m)/.0002
#Partial for Theta 1 & 2 wrt Sigma 1
#start with +eppsilon
mux<-gmu1
muy<-gmu2
muz<-gmu3
sigx<-gsig1+.0001
sigy<-gsig2
sigz<-gsig3
x<-nlminb(start, g, lower = L, upper = U)
o1p<-x$par[1]
o2p<-x$par[2]
#now - eppsilon
mux<-gmu1
muy<-gmu2
muz<-gmu3
sigx<-gsig1-.0001
sigy<-gsig2
sigz<-gsig3
x<-nlminb(start, g, lower = L, upper = U)
o1m<-x$par[1]
o2m<-x$par[2]
#Calc Partial
dc1s1<-(o1p-o1m)/.0002
dc2s1<-(o2p-o2m)/.0002
#Partial for Theta 1 & 2 wrt Sigma 2
#start with +eppsilon
mux<-gmu1
muy<-gmu2
muz<-gmu3
sigx<-gsig1
sigy<-gsig2+.0001
sigz<-gsig3
x<-nlminb(start, g, lower = L, upper = U)
o1p<-x$par[1]
o2p<-x$par[2]
#now - eppsilon
mux<-gmu1
muy<-gmu2
muz<-gmu3
sigx<-gsig1
sigy<-gsig2-.0001
sigz<-gsig3
x<-nlminb(start, g, lower = L, upper = U)
o1m<-x$par[1]
o2m<-x$par[2]
#Calc Partial
dc1s2<-(o1p-o1m)/.0002
dc2s2<-(o2p-o2m)/.0002
#Partial for Theta 1 & 2 wrt Sigma 3
#start with +eppsilon
mux<-gmu1
muy<-gmu2
muz<-gmu3
sigx<-gsig1
```

```
sigy<-gsig2
sigz<-gsig3+.0001
x<-nlminb(start, g, lower = L, upper = U)
o1p<-x$par[1]
o2p<-x$par[2]
#now - eppsilon
mux<-gmu1
muy<-gmu2
muz<-gmu3
sigx<-gsig1
sigy<-gsig2
sigz<-gsig3-.0001
x<-nlminb(start, g, lower = L, upper = U)
o1m<-x$par[1]
o2m<-x$par[2]
#Calc Partial
dc1s3<-(o1p-o1m)/.0002
dc2s3<-(o2p-o2m)/.0002
##calc partial of BC function wrt Mean 1, in three parts
dp1<-(1/gsig1)*((dc2m1-1)*dnorm((c2-gmu1)/gsig1)*(w21*p1-w31*p1)-w21*p1*dnorm((
    c1-gmu1)/gsig1)*(dc1m1-1))
dp2<-(1/gsig2)*(w12*p2*dnorm((c1-gmu2)/gsig2)*dc1m1+w32*p2*dnorm((gmu2-c2)/gsig2
    )*(-dc2m1))
dp3<-(1/gsig3)*((dc1m1)*dnorm((c1-gmu3)/gsig3)*(w13*p3-w23*p3)+w23*p3*dnorm((c2-
    gmu3)/gsig3)*dc2m1)
dbcm1<-dp1+dp2+dp3
##calc partial of BC function wrt Mean 2, in three parts
dp1<-(1/gsig1)*((dc2m2)*dnorm((c2-gmu1)/gsig1)*(w21*p1-w31*p1)-w21*p1*dnorm((c1-
    gmu1)/gsig1)*dc1m2)
dp2<-(1/gsig2)*(w12*p2*dnorm((c1-gmu2)/gsig2)*(dc1m2-1)+w32*p2*dnorm((gmu2-c2)/
    gsig2)*(1-dc2m2))
dp3<-(1/gsig3)*(dc1m2*dnorm((c1-gmu3)/gsig3)*(w13*p3-w23*p3)+w23*p3*dnorm((c2-
    gmu3)/gsig3)*dc2m2)
dbcm2<-dp1+dp2+dp3
##calc partial of BC function wrt Mean 3, in three parts
dp1<-(1/gsig1)*(dc2m3*dnorm((c2-gmu1)/gsig1)*(w21*p1-w31*p1)-w21*p1*dnorm((c1-
    gmu1)/gsig1)*dc1m3)
dp2<-(1/gsig2)*(w12*p2*dnorm((c1-gmu2)/gsig2)*dc1m3+w32*p2*dnorm((gmu2-c2)/gsig2
    )*(-dc2m3))
dp3<-(1/gsig3)*((dc1m3-1)*dnorm((c1-gmu3)/gsig3)*(w13*p3-w23*p3)+w23*p3*dnorm((
    c2-gmu3)/gsig3)*(dc2m3-1))
dbcm3<-dp1+dp2+dp3
##calc partial of BC function wrt Sigma 1, in three parts
dp1<-(1/gsig1)*(dnorm((c2-gmu1)/gsig1)*(w21*p1*(dc2s1-((c2-gmu1)/gsig1))+w31*p1*
    (-dc2s1-((gmu1-c2)/gsig1)))
-w21*p1*dnorm((c1-gmu1)/gsig1)*(dc1s1-((c1-gmu1)/gsig1)))
dp2<-(1/gsig2)*(w12*p2*dnorm((c1-gmu2)/gsig2)*dc1s1-w32*p2*dnorm((gmu2-c2)/gsig2
    )*dc2s1)
dp3<-(1/gsig3)*(dnorm((c1-gmu3)/gsig3)*dc1s1*(w13*p3-w23*p3)+w23*p3*dnorm((c2-
    gmu3)/gsig3)*dc2s1)
dbcs1<-dp1+dp2+dp3
##calc partial of BC function wrt Sigma 2, in three parts
dp1<-(1/gsig1)*(dnorm((c2-gmu1)/gsig1)*dc2s2*(w21*p1-w31*p1)-w21*p1*dnorm((c1-
    gmu1)/gsig1)*dc1s2)
dp2<-(1/gsig2)*(w12*p2*dnorm((c1-gmu2)/gsig2)*(dc1s2-((c1-gmu2)/gsig2))+
  w32*p2*dnorm((gmu2-c2)/gsig2)*(-dc2s2-((gmu2-c2)/gsig2)))
```

```
dp3<-(1/gsig3)*(dnorm((c1-gmu3)/gsig3)*dc1s2*(w13*p3-w23*p3)+w23*p3*dnorm((c2-
    gmu3)/gsig3)*dc2s2)
dbcs2<-dp1+dp2+dp3
##calc partial of BC function wrt Sigma 3, in three parts
dp1<-(1/gsig1)*(dnorm((c2-gmu1)/gsig1)*dc2s3*(w21*p1-w31*p1)-w21*p1*dnorm((c1-
    gmu1)/gsig1)*dc1s3)
dp2<-(1/gsig2)*(w12*p2*dnorm((c1-gmu2)/gsig2)*dc1s3+w32*p2*dnorm((gmu2-c2)/gsig2
    )*(-dc2s3))
dp3<-(1/gsig3)*(dnorm((c1-gmu3)/gsig3)*(dc1s3-((c1-gmu3)/gsig3))*(w13*p3-w23*p3)
    +
  w23*p3*dnorm((c2-gmu3)/gsig3)*(dc2s3-((c2-gmu3)/gsig3)))
dbcs3<-dp1+dp2+dp3
#Calc Variances of Parameters
#var of mean
vm1<-(gsig1^2)/nx
vm2<-(gsig2^2)/ny
vm3<-(gsig3^2)/nz
#var of sigma
vs1<-(gsig1^2)/(2*(nx-1))
vs2<-(gsig2^2)/(2*(ny-1))
vs3<-(gsig3^2)/(2*(nz-1))
#Calc Variance of Bayes Cost
VBC<-(dbcm1^2)*vm1+(dbcs1^2)*vs1+
  (dbcm2^2)*vm2+(dbcs2^2)*vs2+
  (dbcm3^2)*vm3+(dbcs3^2)*vs3
#Calc Variance of Threshold 1
VC1<-(dc1m1^2)*vm1+(dc1s1^2)*vs1+
  (dc1m2^2)*vm2+(dc1s2^2)*vs2+
  (dc1m3^2)*vm3+(dc1s3^2)*vs3
#Calc Variance of Threshold 2
VC2<-(dc2m1^2)*vm1+(dc2s1^2)*vs1+
  (dc2m2^2)*vm2+(dc2s2^2)*vs2+
  (dc2m3^2)*vm3+(dc2s3^2)*vs3
##CI results
LBC1<-c1-1.96*sqrt(VC1)
UBC1<-c1+1.96*sqrt(VC1)
LBC2<-c2-1.96*sqrt(VC2)
UBC2<-c2+1.96*sqrt(VC2)
LBBC<-EBC-1.96*sqrt(VBC)
UBBC<-EBC+1.96*sqrt(VBC)
```

C.1.2 Generalized 95% CIs .

```
#IMPUTS TO CHANGE##
p1<-#SET Prevelance Class 1
p2<-#SET Prevelance Class 2
p3<-#SET Prevelance Class 3
w21<-#SET COST 2|1
w31<-#SET COST 3|1
w12<-#SET COST 1|2
w32<-#SET COST 3|2
w13<-#SET COST 1|3
w23<-#SET COST 2|3
nx<-#SIZE Class 1
ny<-#SIZE Class 2
nz<-#SIZE Class 3
X<-#Vector of Values for Class 1
Y<-#Vector of Values for Class 2
```

```
Z<-#Vector of Values for Class 3
K<-1500 #Change if desire K other than 1500
##Calculations, Do not change
start<-c(-.1,0)
L<-c(-1000,-1000)
U<-c(1000,1000)
ybar1<-mean(X)
ybar2<-mean(Y)
ybar3<-mean(Z)
var1<-var(X)
var2<-var(Y)
var3<-var(Z)
##Create Pivotal Quantile for each Mean, and Var
t1<-rt(K,nx-1)
t2<-rt(K,ny-1)
t3<-rt(K,nz-1)
V1<-rchisq(K,nx-1)
V2<-rchisq(K,ny-1)
V3<-rchisq(K,nz-1)
Rs1<-c(rep((nx-1)*var1,K))/V1
Rs2<-c(rep((ny-1)*var2,K))/V2
Rs3<-c(rep((nz-1)*var3,K))/V3
Rm1<-c(rep(ybar1,K))-(t1*(sqrt(var1/nx)))
Rm2<-c(rep(ybar2,K))-(t2*(sqrt(var2/ny)))
Rm3<-c(rep(ybar3,K))-(t3*(sqrt(var3/nz)))
#Find K BC and Opt. Threshold values using Numerical Minimization
BC<-c(rep(-9999,K))
C1<-c(rep(-9999,K))
C2<-c(rep(-9999,K))
for (i in 1:K){
h<-function(par){(pnorm(par[2],Rm1[i],sqrt(Rs1[i]))-pnorm(par[1],Rm1[i],sqrt(Rs1
    [i])))*(p1*w21)+
(1-pnorm(par[2],Rm1[i],sqrt(Rs1[i])))*(p1*w31)+
(pnorm(par[1],Rm2[i],sqrt(Rs2[i])))*(p2*w12)+
(1-pnorm(par[2],Rm2[i],sqrt(Rs2[i])))*(p2*w32)+
(pnorm(par[1],Rm3[i],sqrt(Rs3[i])))*(p3*w13)+
(pnorm(par[2],Rm3[i],sqrt(Rs3[i]))-pnorm(par[1],Rm3[i],sqrt(Rs3[i])))*(p3*w23)}
sols<-optim(start, h, lower = L, upper = U, method="L-BFGS-B")
BC[i]<-sols$value
C1[i]<-sols$par[1]
C2[i]<-sols$par[2]
}
# CI Results
LBC1<-quantile(C1,.025)
UBC1<-quantile(C1,.975)
LBC2<-quantile(C2,.025)
UBC2<-quantile(C2,.975)
LBBC<-quantile(BC,.025)
UBBC<-quantile(BC,.975)
```

C.1.3 Fiducial 95% CI for BC with Equal Weights.

```
##INPUTS for Setup##
n1<-#Sample Size Class 1
n2<-#Sample Size Class 1
n3<-#Sample Size Class 1
BChat<-#Estimated BC
###Do not Change
```

```
g<-c(1,0,1,0,1,0)
Umat<-c(1/n1,0,1/n2,0,1/n3,0)
BChat<-round(BChat,5)
row=(n1+1)*2
ss=matrix(seq(from=0,to=row-1,by=1), ncol=2)
ss[,2]=n1-ss[,1]
ss1<-ss[,1:2]
row=(n2+1)*2
ss=matrix(seq(from=0,to=row-1,by=1), ncol=2)
ss[,2]=n2-ss[,1]
ss2<-ss[,1:2]
row=(n3+1)*2
ss=matrix(seq(from=0,to=row-1,by=1), ncol=2)
ss[,2]=n3-ss[,1]
ss3<-ss[,1:2]
LEN<-length(ss1[,1])*length(ss2[,1])*length(ss3[,1])
len1<-length(ss1[,1])
len2<-length(ss2[,1])
len3<-length(ss3[,1])
v1<-c(rep(1,len2*len3))
col1<-kronecker(ss1,v1)
v2<-c(rep(1,len1))
v3<-c(rep(1,len3))
col2<-kronecker(v2,kronecker(ss2,v3))
v4<-c(rep(1,len1*len2))
col3<-kronecker(v4,ss3)
SS<-matrix(cbind(col1,col2,col3),ncol=6)
U1<-SS[,1:6]%*%Umat
U1<-round(U1,5)
SS<-cbind(SS,U1)
temp<-data.frame(SS)
SSOR<-temp[order(temp[,7]),]
##CREATE Probability SAMPLE SPACES
##by .1, SSP1
pvec<-seq(from=0, to=1, by=.1)
len<-length(pvec)
v1<-c(rep(1,len*len))
col1<-kronecker(pvec,v1)
v2<-c(rep(1,len))
col2<-kronecker(v2,kronecker(pvec,v2))
col3<-kronecker(v2,kronecker(v2,pvec))
c12<-c(1-col1)
c22<-c(1-col2)
c32<-c(1-col3)
SSP1<-cbind(col1,c12,col2,c22,col3,c32)
##by .05, SSP2
pvec<-seq(from=0, to=1, by=.05)
len<-length(pvec)
v1<-c(rep(1,len*len))
col1<-kronecker(pvec,v1)
v2<-c(rep(1,len))
col2<-kronecker(v2,kronecker(pvec,v2))
col3<-kronecker(v2,kronecker(v2,pvec))
c12<-c(1-col1)
c22<-c(1-col2)
c32<-c(1-col3)
SSP2<-cbind(col1,c12,col2,c22,col3,c32)
```

```
##by .01, SSP3
pvec<-seq(from=0, to=1, by=.01)
len<-length(pvec)
v1<-c(rep(1,len*len))
col1<-kronecker(pvec,v1)
v2<-c(rep(1,len))
col2<-kronecker(v2,kronecker(pvec,v2))
col3<-kronecker(v2,kronecker(v2,pvec))
c12<-c(1-col1)
c22<-c(1-col2)
c32<-c(1-col3)
SSP3<-cbind(col1,c12,col2,c22,col3,c32)
##by .005, SSP4
pvec<-seq(from=0, to=1, by=.005)
len<-length(pvec)
v1<-c(rep(1,len*len))
col1<-kronecker(pvec,v1)
v2<-c(rep(1,len))
col2<-kronecker(v2,kronecker(pvec,v2))
col3<-kronecker(v2,kronecker(v2,pvec))
c12<-c(1-col1)
c22<-c(1-col2)
c32<-c(1-col3)
SSP4<-cbind(col1,c12,col2,c22,col3,c32)
end<-length(SSOR[,1])
##Define Partial CDFS
f1<-function(p){
factorial(n1)*factorial(n2)*factorial(n3)*sum(((p[1]^(SSOR[(UBound+1):end,1]))/
    factorial(SSOR[(UBound+1):end,1]))*((p[2]^(SSOR[(UBound+1):end,2]))/
    factorial(SSOR[(UBound+1):end,2]))*
((p[3]^(SSOR[(UBound+1):end,3]))/factorial(SSOR[(UBound+1):end,3]))*((p[4]^(SSOR
    [(UBound+1):end,4]))/factorial(SSOR[(UBound+1):end,4]))*
((p[5]^(SSOR[(UBound+1):end,5]))/factorial(SSOR[(UBound+1):end,5]))*((p[6]^(SSOR
    [(UBound+1):end,6]))/factorial(SSOR[(UBound+1):end,6])))
}
f2<-function(p){
factorial(n1)*factorial(n2)*factorial(n3)*sum(((p[1]^(SSOR[LBound:UBound,1]))/
    factorial(SSOR[LBound:UBound,1]))*((p[2]^(SSOR[LBound:UBound,2]))/factorial(
    SSOR[LBound:UBound,2]))*
((p[3]^(SSOR[LBound:UBound,3]))/factorial(SSOR[LBound:UBound,3]))*((p[4]^(SSOR[
    LBound:UBound,4]))/factorial(SSOR[LBound:UBound,4]))*
((p[5]^(SSOR[LBound:UBound,5]))/factorial(SSOR[LBound:UBound,5]))*((p[6]^(SSOR[
    LBound:UBound,6]))/factorial(SSOR[LBound:UBound,6])))
}
BCmatch<-which(SSOR[,7]==BChat[1])
LBound<-min(BCmatch)
UBound<-max(BCmatch)
##Find Solution 1st Iteration####
BCOUT1<-apply(SSP1, 1, FUN = f1)
BCOUT2<-apply(SSP1, 1, FUN = f2)
BCL<-BCOUT1+BCOUT2
BCU<-c(rep(1,length(BCOUT1)))-BCOUT1
BC<-SSP1%*%g
BC<-round(BC,5)
blah2<-cbind(SSP1,BCU,BC)
BCcdf<-unique(BC)
BCcdf<-cbind(BCcdf,rep(-999,length(BCcdf)))
```

```
for (i in 1:length(BCcdf[,1])){
BCcdf[i,2]<-max(blah2[which(blah2[,8]==BCcdf[i,1]),7])
}
#GET UB
temp<-max(BCcdf[which(BCcdf[,2]<=0.025),2])
UB<-min(BCcdf[which(BCcdf[,2]==temp),1])
#GET LB
BC<-SSP1%*%g
BC<-round(BC,5)
blah2<-cbind(SSP1,BCL,BC)
BCcdf<-unique(BC)
BCcdf<-cbind(BCcdf,rep(-999,length(BCcdf)))
for (i in 1:length(BCcdf[,1])){
BCcdf[i,2]<-max(blah2[which(blah2[,8]==BCcdf[i,1]),7])
}
temp<-max(BCcdf[which(BCcdf[,2]<=0.025),2])
LB<-max(BCcdf[which(BCcdf[,2]==temp),1])
##Refine in on Solution 1st time
BC<-SSP2%*%g
BC<-round(BC,5)
SSPtemp<-cbind(SSP2,BC)
SSPn<-SSP2[which(SSPtemp[,7]<=(UB+.2)&SSPtemp[,7]>=(UB-.2)),]
BCOUT1<-apply(SSPn, 1, FUN = f1)
BCU<-c(rep(1,length(BCOUT1)))-BCOUT1
BC<-SSPn%*%g
BC<-round(BC,5)
blah2<-cbind(SSPn,BCU,BC)
BCcdf<-unique(BC)
BCcdf<-cbind(BCcdf,rep(-999,length(BCcdf)))
for (i in 1:length(BCcdf[,1])){
BCcdf[i,2]<-max(blah2[which(blah2[,8]==BCcdf[i,1]),7])
}
#GET UB
temp<-max(BCcdf[which(BCcdf[,2]<=0.025),2])
UB<-min(BCcdf[which(BCcdf[,2]==temp),1])
#GET LB
BC<-SSP2%*%g
BC<-round(BC,5)
SSPtemp<-cbind(SSP2,BC)
SSPn<-SSP2[which(SSPtemp[,7]<=(LB+.2)&SSPtemp[,7]>=(LB-.2)),]
BCOUT1<-apply(SSPn, 1, FUN = f1)
BCOUT2<-apply(SSPn, 1, FUN = f2)
BCL<-BCOUT1+BCOUT2
BC<-SSPn%*%g
BC<-round(BC,5)
blah2<-cbind(SSPn,BCL,BC)
BCcdf<-unique(BC)
BCcdf<-cbind(BCcdf,rep(-999,length(BCcdf)))
for (i in 1:length(BCcdf[,1])){
BCcdf[i,2]<-max(blah2[which(blah2[,8]==BCcdf[i,1]),7])
}
temp<-max(BCcdf[which(BCcdf[,2]<=0.025),2])
LB<-max(BCcdf[which(BCcdf[,2]==temp),1])
##Refine in on Solution 2nd time
BC<-SSP3%*%g
BC<-round(BC,5)
SSPtemp<-cbind(SSP3,BC)
```

```
SSPn<-SSP3[which(SSPtemp[,7]<=(UB+.1)&SSPtemp[,7]>=(UB-.1)),]
BCOUT1<-apply(SSPn, 1, FUN = f1)
BCU<-c(rep(1,length(BCOUT1)))-BCOUT1
BC<-SSPn%*%g
BC<-round(BC,5)
blah2<-cbind(SSPn,BCU,BC)
BCcdf<-unique(BC)
BCcdf<-cbind(BCcdf,rep(-999,length(BCcdf)))
for (i in 1:length(BCcdf[,1])){
BCcdf[i,2]<-max(blah2[which(blah2[,8]==BCcdf[i,1]),7])
}
#GET UB
temp<-max(BCcdf[which(BCcdf[,2]<=0.025),2])
UB<-min(BCcdf[which(BCcdf[,2]==temp),1])
#GET LB
BC<-SSP3%*%g
BC<-round(BC,5)
SSPtemp<-cbind(SSP3,BC)
SSPn<-SSP3[which(SSPtemp[,7]<=(LB+.1)&SSPtemp[,7]>=(LB-.1)),]
BCOUT1<-apply(SSPn, 1, FUN = f1)
BCOUT2<-apply(SSPn, 1, FUN = f2)
BCL<-BCOUT1+BCOUT2
BC<-SSPn%*%g
BC<-round(BC,5)
blah2<-cbind(SSPn,BCL,BC)
BCcdf<-unique(BC)
BCcdf<-cbind(BCcdf,rep(-999,length(BCcdf)))
for (i in 1:length(BCcdf[,1])){
BCcdf[i,2]<-max(blah2[which(blah2[,8]==BCcdf[i,1]),7])
}
temp<-max(BCcdf[which(BCcdf[,2]<=0.025),2])
LB<-max(BCcdf[which(BCcdf[,2]==temp),1])
##Refine in on Solution 3rd time
BC<-SSP4%*%g
BC<-round(BC,5)
SSPtemp<-cbind(SSP4,BC)
SSPn<-SSP4[which(SSPtemp[,7]<=(UB+.05)&SSPtemp[,7]>=(UB-.05)),]
BCOUT1<-apply(SSPn, 1, FUN = f1)
BCU<-c(rep(1,length(BCOUT1)))-BCOUT1
BC<-SSPn%*%g
BC<-round(BC,5)
blah2<-cbind(SSPn,BCU,BC)
BCcdf<-unique(BC)
BCcdf<-cbind(BCcdf,rep(-999,length(BCcdf)))
for (i in 1:length(BCcdf[,1])){
BCcdf[i,2]<-max(blah2[which(blah2[,8]==BCcdf[i,1]),7])
}
#GET UB
temp<-max(BCcdf[which(BCcdf[,2]<=0.025),2])
UB<-min(BCcdf[which(BCcdf[,2]==temp),1])
#GET LB
BC<-SSP4%*%g
BC<-round(BC,5)
SSPtemp<-cbind(SSP4,BC)
SSPn<-SSP4[which(SSPtemp[,7]<=(LB+.05)&SSPtemp[,7]>=(LB-.05)),]
BCOUT1<-apply(SSPn, 1, FUN = f1)
BCOUT2<-apply(SSPn, 1, FUN = f2)
```

```
BCL<-BCOUT1+BCOUT2
BC<-SSPn%*%g
BC<-round(BC,5)
blah2<-cbind(SSPn,BCL,BC)
BCcdf<-unique(BC)
BCcdf<-cbind(BCcdf,rep(-999,length(BCcdf)))
for (i in 1:length(BCcdf[,1])){
BCcdf[i,2]<-max(blah2[which(blah2[,8]==BCcdf[i,1]),7])
}
templ<-max(BCcdf[which(BCcdf[,2]<=0.025),2])
LB<-max(BCcdf[which(BCcdf[,2]==templ),1])
#CI Results
print(c(LB,UB))
```

C.1.4 Fiducial 95% CI for BC with Unequal Weights.

```
#Inputs to Change
p1<-#SET Prevelance Class 1
p2<-#SET Prevelance Class 2
p3<-#SET Prevelance Class 3
w21<-#SET COST 2|1
w31<-#SET COST 3|1
w12<-#SET COST 1|2
w32<-#SET COST 3|2
w13<-#SET COST 1|3
w23<-#SET COST 2|3
n1<-#Sample Size Class 1
n2<-#Sample Size Class 2
n3<-#Sample Size Class 3
BChat<-#Estimated Bayes Cost
##CREATE MULTINOMIAL SAMPLESPACE VIA Weizhen Wang 2012
#Class 1 SS
row=(n1+1)*(n1+2)/2*4
ss=matrix(1:row, ncol=4)
nn=1
fn=1
while (nn<n+1+0.5){
low=fn-nn+1
up=fn
ss[low:up,1]=n+1-nn
uu=up-low
ss[low:up,2]=0:uu
nn=nn+1
fn=fn+nn
}
ss[,3]=n-ss[,1]-ss[,2]
ss1<-ss[,1:3]
#Class 2 SS
row=(n2+1)*(n2+2)/2*4
ss=matrix(1:row, ncol=4)
nn=1
fn=1
while (nn<n+1+0.5){
low=fn-nn+1
up=fn
ss[low:up,1]=n+1-nn
uu=up-low
ss[low:up,2]=0:uu
```

```
nn=nn+1
fn=fn+nn
}
ss[,3]=n-ss[,1]-ss[,2]
ss2<-ss[,1:3]
#Class 3 SS
row=(n3+1)*(n3+2)/2*4
ss=matrix(1:row, ncol=4)
nn=1
fn=1
while (nn<n+1+0.5){
low=fn-nn+1
up=fn
ss[low:up,1]=n+1-nn
uu=up-low
ss[low:up,2]=0:uu
nn=nn+1
fn=fn+nn
}
ss[,3]=n-ss[,1]-ss[,2]
ss3<-ss[,1:3]
LEN<-length(ss1[,1])*length(ss2[,1])*length(ss3[,1])
len1<-length(ss1[,1])
len2<-length(ss2[,1])
len3<-length(ss3[,1])
v1<-c(rep(1,len2*len3))
col1<-kronecker(ss1,v1)
v2<-c(rep(1,len1))
v3<-c(rep(1,len3))
col2<-kronecker(v2,kronecker(ss2,v3))
v4<-c(rep(1,len1*len2))
col3<-kronecker(v4,ss3)
SS<-matrix(cbind(col1,col2,col3),ncol=9)
Umat<-c(0,(p1*w21)/n1,(p1*w31)/n1,0,(p2*w12)/n2,(p2*w32)/n2,0,(p3*w13)/n3,(p3*
    w23)/n3)
U1<-SS[,1:9]%*%Umat
U1<-round(U1,5)
SS<-cbind(SS,U1)
##Order BC sample Space
temp<-data.frame(SS)
SSOR<-temp[order(temp[,10]),]
end<-length(SSOR[,1])
##by .05, SSP3
pvec<-seq(from=0, to=1, by=.05)
len<-length(pvec)
v1<-c(rep(1,len))
col1<-kronecker(pvec,v1)
col2<-kronecker(v1,pvec)
col3<-1-col1-col2
Ps<-cbind(col1,col2,col3)
Pspace<-Ps[-which(Ps[,3]<0),]
rowp<-length(Pspace[,1])
v1<-c(rep(1,rowp*rowp))
col1<-kronecker(Pspace,v1)
v2<-c(rep(1,rowp))
col2<-kronecker(v2,kronecker(Pspace,v2))
col3<-kronecker(v2,kronecker(v2,Pspace))
```

```
SSP1<-matrix(cbind(col1,col2,col3),ncol=9)
g<-c(0,(p1*w21),(p1*w31),0,(p2*w12),(p2*w32),0,(p3*w13),(p3*w23))
f1<-function(p){
factorial(n1)*factorial(n2)*factorial(n3)*sum(((p[1]^(SSOR[1:(LBound-1),1]))/
    factorial(SSOR[1:(LBound-1),1]))*((p[2]^(SSOR[1:(LBound-1),2]))/factorial(
    SSOR[1:(LBound-1),2]))*
((p[3]^(SSOR[1:(LBound-1),3]))/factorial(SSOR[1:(LBound-1),3]))*((p[4]^(SSOR[1:(
    LBound-1),4]))/factorial(SSOR[1:(LBound-1),4]))*
((p[5]^(SSOR[1:(LBound-1),5]))/factorial(SSOR[1:(LBound-1),5]))*((p[6]^(SSOR[1:(
    LBound-1),6]))/factorial(SSOR[1:(LBound-1),6]))*
((p[7]^(SSOR[1:(LBound-1),7]))/factorial(SSOR[1:(LBound-1),7]))*((p[8]^(SSOR[1:(
    LBound-1),8]))/factorial(SSOR[1:(LBound-1),8]))*
((p[9]^(SSOR[1:(LBound-1),9]))/factorial(SSOR[1:(LBound-1),9])))
}
f2<-function(p){
factorial(n1)*factorial(n2)*factorial(n3)*sum(((p[1]^(SSOR[LBound:UBound,1]))/
    factorial(SSOR[LBound:UBound,1]))*((p[2]^(SSOR[LBound:UBound,2]))/factorial(
    SSOR[LBound:UBound,2]))*
((p[3]^(SSOR[LBound:UBound,3]))/factorial(SSOR[LBound:UBound,3]))*((p[4]^(SSOR[
    LBound:UBound,4]))/factorial(SSOR[LBound:UBound,4]))*
((p[5]^(SSOR[LBound:UBound,5]))/factorial(SSOR[LBound:UBound,5]))*((p[6]^(SSOR[
    LBound:UBound,6]))/factorial(SSOR[LBound:UBound,6]))*
((p[7]^(SSOR[LBound:UBound,7]))/factorial(SSOR[LBound:UBound,7]))*((p[8]^(SSOR[
    LBound:UBound,8]))/factorial(SSOR[LBound:UBound,8]))*
((p[9]^(SSOR[LBound:UBound,9]))/factorial(SSOR[LBound:UBound,9])))
}
BCmatch<-which(SSOR[,10]==BChat[1])
LBound<-min(BCmatch)
UBound<-max(BCmatch)
BCOUT1<-apply(SSP1, 1, FUN = f1)
BCOUT2<-apply(SSP1, 1, FUN = f2)
BCU<-BCOUT1+BCOUT2
BCL<-c(rep(1,length(BCOUT1)))-BCOUT1
BC<-SSP1%*%g
BC<-round(BC,5)
blah2<-cbind(SSP1,BCU,BC)
BCcdf<-unique(BC)
BCcdf<-cbind(BCcdf,rep(-999,length(BCcdf)))
for (i in 1:length(BCcdf[,1])){
BCcdf[i,2]<-max(blah2[which(blah2[,11]==BCcdf[i,1]),10])
}
BCcdf<-data.frame(BCcdf)
BCcdf<-BCcdf[order(BCcdf[,1]),]
BCcdfs<-BCcdf
blah<-length(BCcdfs[,1])
for (i in 1:blah){
BCcdfs[i,2]<-max(BCcdf[which(BCcdf[,1]>=BCcdfs[i,1]),2])
}
BCcdf<-BCcdfs
#GET UB
temp<-max(BCcdf[which(BCcdf[,2]<=0.025),2])
UB<-min(BCcdf[which(BCcdf[,2]==temp),1])
#GET LB
BC<-SSP1%*%g
BC<-round(BC,5)
blah2<-cbind(SSP1,BCL,BC)
BCcdf<-unique(BC)
```

```
BCcdf<-cbind(BCcdf,rep(-999,length(BCcdf)))
for (i in 1:length(BCcdf[,1])){
BCcdf[i,2]<-max(blah2[which(blah2[,11]==BCcdf[i,1]),10])
}
BCcdf<-data.frame(BCcdf)
BCcdf<-BCcdf[order(BCcdf[,1]),]
BCcdfs<-BCcdf
blah<-length(BCcdfs[,1])
for (i in 1:blah){
BCcdfs[i,2]<-max(BCcdf[which(BCcdf[,1]<=BCcdfs[i,1]),2])
}
BCcdf<-BCcdfs
temp<-max(BCcdf[which(BCcdf[,2]<=0.025),2])
LB<-max(BCcdf[which(BCcdf[,2]==temp),1])
#CI Results
print(c(LB,UB))
```

C.1.5 Delta Method Hypothesis Tests .

C.1.5.1 One-Sided Test on Single BC Value.

```
#Set Up
p1<-#SET Prevelance Class 1
p2<-#SET Prevelance Class 2
p3<-#SET Prevelance Class 3
w21<-#SET COST 2|1
w31<-#SET COST 3|1
w12<-#SET COST 1|2
w32<-#SET COST 3|2
w13<-#SET COST 1|3
w23<-#SET COST 2|3
TV<-# Set BCnot
n1<-#SIZE Class 1
n2<-#SIZE Class 2
n3<-#SIZE Class 3
Y<-#Vector of Values for Class 1
X<-#Vector of Values for Class 2
Z<-#Vector of Values for Class 3
start<-c(-.1,0)
L<-c(-1000,-1000)
U<-c(1000,1000)
##Do Not Change
gmu1<-mean(Y)
gmu2<-mean(X)
gmu3<-mean(Z)
gsig1<-sd(Y)
gsig2<-sd(X)
gsig3<-sd(Z)
f<-function(par){abs(pnorm(par[2],gmu1,gsig1)-pnorm(par[1],gmu1,gsig1))*(p1*w21)
    +
abs(1-pnorm(par[2],gmu1,gsig1))*(p1*w31)+
abs(pnorm(par[1],gmu2,gsig2))*(p2*w12)+
abs(1-pnorm(par[2],gmu2,gsig2))*(p2*w32)+
abs(pnorm(par[1],gmu3,gsig3))*(p3*w13)+
abs(pnorm(par[2],gmu3,gsig3)-pnorm(par[1],gmu3,gsig3))*(p3*w23)}
x<-nlminb(start, f, lower = L, upper = U)
c1<-x$par[1]
c2<-x$par[2]
EBC<-x$objective
```

```
#Calculate All Partial Derivatives as was done for Delta Method CI#
#Calc Variances of Parameters
vm1<-(gsig1^2)/n1
vm2<-(gsig2^2)/n2
vm3<-(gsig3^2)/n3
#Calc var of sig using delta method
vs1<-(gsig1^2)/(2*(n1-1))
vs2<-(gsig2^2)/(2*(n1-1))
vs3<-(gsig3^2)/(2*(n1-1))
VBC<-(dbcm1^2)*vm1+(dbcs1^2)*vs1+
  (dbcm2^2)*vm2+(dbcs2^2)*vs2+
  (dbcm3^2)*vm3+(dbcs3^2)*vs3
W<-(EBC-TV)/sqrt(VBC)
#Test p-value - to compare to alpha
deltap<-pnorm(W, lower.tail=TRUE)
```

C.1.5.2 One-Sided Test on the Difference of Two Independent BC Values.

```
p1<-#SET Prevelance Class 1
p2<-#SET Prevelance Class 2
p3<-#SET Prevelance Class 3
w21<-#SET COST 2|1
w31<-#SET COST 3|1
w12<-#SET COST 1|2
w32<-#SET COST 3|2
w13<-#SET COST 1|3
w23<-#SET COST 2|3
TV<-#Set BCnot
n1<-#SIZE Class 1
n2<-#SIZE Class 2
n3<-#SIZE Class 3
TV<-#Set Eta_not
YA<-#Vector of Values for Class 1-Classification System A
XA<-#Vector of Values for Class 2-Classification System A
ZA<-#Vector of Values for Class 3-Classification System A
Y<-#Vector of Values for Class 1-Classification System B
X<-#Vector of Values for Class 2-Classification System B
Z<-#Vector of Values for Class 3-Classification System B
##Do Not Change
#CS A
gmu1<-mean(YA)
gmu2<-mean(XA)
gmu3<-mean(ZA)
gsig1<-sd(YA)
gsig2<-sd(XA)
gsig3<-sd(ZA)
f<-function(par){abs(pnorm(par[2],gmu1,gsig1)-pnorm(par[1],gmu1,gsig1))*(p1*w21)
    +
abs(1-pnorm(par[2],gmu1,gsig1))*(p1*w31)+
abs(pnorm(par[1],gmu2,gsig2))*(p2*w12)+
abs(1-pnorm(par[2],gmu2,gsig2))*(p2*w32)+
abs(pnorm(par[1],gmu3,gsig3))*(p3*w13)+
abs(pnorm(par[2],gmu3,gsig3)-pnorm(par[1],gmu3,gsig3))*(p3*w23)}
x<-optim(start, f, lower = L, upper = U,method="L-BFGS-B")
c1<-x$par[1]
c2<-x$par[2]
EBCA<-x$value
#Calculate all Partial Derivatives for CS A as was done for Delta CI
```

```
#Calc Variances of Parameters
vm1<-(gsig1^2)/n1
vm2<-(gsig2^2)/n2
vm3<-(gsig3^2)/n3
#Calc var of sig using delta method
vs1<-(gsig1^2)/(2*(n1-1))
vs2<-(gsig2^2)/(2*(n1-1))
vs3<-(gsig3^2)/(2*(n1-1))
VBCA<-(dbcm1^2)*vm1+(dbcs1^2)*vs1+
  (dbcm2^2)*vm2+(dbcs2^2)*vs2+
  (dbcm3^2)*vm3+(dbcs3^2)*vs3
#CS B
gmu1<-mean(Y)
gmu2<-mean(X)
gmu3<-mean(Z)
gsig1<-sd(Y)
gsig2<-sd(X)
gsig3<-sd(Z)
f<-function(par){abs(pnorm(par[2],gmu1,gsig1)-pnorm(par[1],gmu1,gsig1))*(p1*w21)
    +
abs(1-pnorm(par[2],gmu1,gsig1))*(p1*w31)+
abs(pnorm(par[1],gmu2,gsig2))*(p2*w12)+
abs(1-pnorm(par[2],gmu2,gsig2))*(p2*w32)+
abs(pnorm(par[1],gmu3,gsig3))*(p3*w13)+
abs(pnorm(par[2],gmu3,gsig3)-pnorm(par[1],gmu3,gsig3))*(p3*w23)}
x<-optim(start, f, lower = L, upper = U,method="L-BFGS-B")
c1<-x$par[1]
c2<-x$par[2]
EBC<-x$value
#Calculate all Partial Derivatives for CS B as was done for Delta CI
#Calc Variances of Parameters
vm1<-(gsig1^2)/n1
vm2<-(gsig2^2)/n2
vm3<-(gsig3^2)/n3
#Calc var of sig using delta method
vs1<-(gsig1^2)/(2*(n1-1))
vs2<-(gsig2^2)/(2*(n1-1))
vs3<-(gsig3^2)/(2*(n1-1))
VBC<-(dbcm1^2)*vm1+(dbcs1^2)*vs1+
  (dbcm2^2)*vm2+(dbcs2^2)*vs2+
  (dbcm3^2)*vm3+(dbcs3^2)*vs3
VETA<-VBCA+VBC
EETA<-EBCA-EBC
W<-(EETA-TV)/sqrt(VETA)
#Test p-value - to compare to alpha
deltap<-pnorm(W, lower.tail=FALSE)
```

C.1.6 Generalized Hypothesis Tests .

C.1.6.1 One-Sided Test on Single BC Value.

```
#Set Up
p1<-#SET Prevelance Class 1
p2<-#SET Prevelance Class 2
p3<-#SET Prevelance Class 3
w21<-#SET COST 2|1
w31<-#SET COST 3|1
w12<-#SET COST 1|2
w32<-#SET COST 3|2
```

```
w13<-#SET COST 1|3
w23<-#SET COST 2|3
TV<-#Set BCnot
n1<-#SIZE Class 1
n2<-#SIZE Class 2
n3<-#SIZE Class 3
Y<-#Vector of Values for Class 1
X<-#Vector of Values for Class 2
Z<-#Vector of Values for Class 3
K<-2500 #Change if K other than 2500 is desired
start<-c(-.1,0)
L<-c(-1000,-1000)
U<-c(1000,1000)
##Do Not Change
ybar2<-mean(Y)
ybar1<-mean(X)
ybar3<-mean(Z)
var2<-var(Y)
var1<-var(X)
var3<-var(Z)
t1<-rt(K,n2-1)
t2<-rt(K,n1-1)
t3<-rt(K,n3-1)
V1<-rchisq(K,n2-1)
V2<-rchisq(K,n1-1)
V3<-rchisq(K,n3-1)
Rs1<-c(rep((n2-1)*var1,K))/V1
Rs2<-c(rep((n1-1)*var2,K))/V2
Rs3<-c(rep((n3-1)*var3,K))/V3
Rm1<-c(rep(ybar1,K))-(t1*(sqrt(var1/n2)))
Rm2<-c(rep(ybar2,K))-(t2*(sqrt(var2/n1)))
Rm3<-c(rep(ybar3,K))-(t3*(sqrt(var3/n3)))
f<-function(x){
hun2<-function(par){abs(pnorm(par[2],x[1],x[2])-pnorm(par[1],x[1],x[2]))*(p1*w21
    )+
abs(1-pnorm(par[2],x[1],x[2]))*(p1*w31)+
abs(pnorm(par[1],x[3],x[4]))*(p2*w12)+
abs(1-pnorm(par[2],x[3],x[4]))*(p2*w32)+
abs(pnorm(par[1],x[5],x[6]))*(p3*w13)+
abs(pnorm(par[2],x[5],x[6])-pnorm(par[1],x[5],x[6]))*(p3*w23)}
y<-optim(start, hun2, lower = L, upper = U,method="L-BFGS-B")
BC<-y$value
return(BC)
}
ap1<-cbind(Rm2,sqrt(Rs2),Rm1,sqrt(Rs1),Rm3,sqrt(Rs3))
RBC<-apply(ap1, 1, FUN=f)
#Test p-value - to compare to alpha
genp<-length(which(RBC>TV))/K
```

C.1.6.2 One-Sided Test on the Difference of Two Independent BC Values.

```
p1<-#SET Prevelance Class 1
p2<-#SET Prevelance Class 2
p3<-#SET Prevelance Class 3
w21<-#SET COST 2|1
w31<-#SET COST 3|1
w12<-#SET COST 1|2
w32<-#SET COST 3|2
```

```
w13<-#SET COST 1|3
w23<-#SET COST 2|3
TV<-#Set BCnot
n1<-#SIZE Class 1
n2<-#SIZE Class 2
n3<-#SIZE Class 3
TV<-#Set Eta_not
YA<-#Vector of Values for Class 1-Classification System A
XA<-#Vector of Values for Class 2-Classification System A
ZA<-#Vector of Values for Class 3-Classification System A
Y<-#Vector of Values for Class 1-Classification System B
X<-#Vector of Values for Class 2-Classification System B
Z<-#Vector of Values for Class 3-Classification System B
K<-2500 # Change if desire K other than 2500
##Do Not Change
ybar2<-mean(YA)
ybar1<-mean(XA)
ybar3<-mean(ZA)
var2<-var(YA)
var1<-var(XA)
var3<-var(ZA)
t1<-rt(K,n2-1)
t2<-rt(K,n1-1)
t3<-rt(K,n3-1)
V1<-rchisq(K,n2-1)
V2<-rchisq(K,n1-1)
V3<-rchisq(K,n3-1)
Rs1<-c(rep((n2-1)*var1,K))/V1
Rs2<-c(rep((n1-1)*var2,K))/V2
Rs3<-c(rep((n3-1)*var3,K))/V3
Rm1<-c(rep(ybar1,K))-(t1*(sqrt(var1/n2)))
Rm2<-c(rep(ybar2,K))-(t2*(sqrt(var2/n1)))
Rm3<-c(rep(ybar3,K))-(t3*(sqrt(var3/n3)))
f<-function(x){
hun2<-function(par){abs(pnorm(par[2],x[1],x[2])-pnorm(par[1],x[1],x[2]))*(p1*w21
    )+
abs(1-pnorm(par[2],x[1],x[2]))*(p1*w31)+
abs(pnorm(par[1],x[3],x[4]))*(p2*w12)+
abs(1-pnorm(par[2],x[3],x[4]))*(p2*w32)+
abs(pnorm(par[1],x[5],x[6]))*(p3*w13)+
abs(pnorm(par[2],x[5],x[6])-pnorm(par[1],x[5],x[6]))*(p3*w23)}
y<-optim(start, hun2, lower = L, upper = U,method="L-BFGS-B")
BC<-y$value
return(BC)
}
ap1<-cbind(Rm2,sqrt(Rs2),Rm1,sqrt(Rs1),Rm3,sqrt(Rs3))
RbcA<-apply(ap1, 1, FUN=f)
ybar2<-mean(Y)
ybar1<-mean(X)
ybar3<-mean(Z)
var2<-var(Y)
var1<-var(X)
var3<-var(Z)
t1<-rt(K,n2-1)
t2<-rt(K,n1-1)
t3<-rt(K,n3-1)
V1<-rchisq(K,n2-1)
```

```
V2<-rchisq(K,n1-1)
V3<-rchisq(K,n3-1)
Rs1<-c(rep((n2-1)*var1,K))/V1
Rs2<-c(rep((n1-1)*var2,K))/V2
Rs3<-c(rep((n3-1)*var3,K))/V3
Rm1<-c(rep(ybar1,K))-(t1*(sqrt(var1/n2)))
Rm2<-c(rep(ybar2,K))-(t2*(sqrt(var2/n1)))
Rm3<-c(rep(ybar3,K))-(t3*(sqrt(var3/n3)))
ap1<-cbind(Rm2,sqrt(Rs2),Rm1,sqrt(Rs1),Rm3,sqrt(Rs3))
Rbc<-apply(ap1, 1, FUN=f)
Reta<-RbcA-Rbc
#Test p-value - to compare to alpha
genp<-length(which(Reta<TV))/K
```

C.1.7 Exact Hypothesis Tests .

C.1.7.1 One-Sided Test on Single BC Value.

```
#Inputs
BC0<-#set BC_not
n1<-#Sample Size Class 1
n2<-#Sample Size Class 1
n3<-#Sample Size Class 1
BChat<-#Estimated BC
p1<-#SET Prevelance Class 1
p2<-#SET Prevelance Class 2
p3<-#SET Prevelance Class 3
w21<-#SET COST 2|1
w31<-#SET COST 3|1
w12<-#SET COST 1|2
w32<-#SET COST 3|2
w13<-#SET COST 1|3
w23<-#SET COST 2|3
##Creat SSOR as done in Fiducial CI code
#Create Probability Space
pvec<-seq(from=0, to=1, by=.05)
len<-length(pvec)
v1<-c(rep(1,len))
col1<-kronecker(pvec,v1)
col2<-kronecker(v1,pvec)
col3<-1-col1-col2
Ps<-cbind(col1,col2,col3)
Pspace3<-Ps[-which(Ps[,3]<0),]
col1b<-pvec
c12<-1-pvec
Pspace2<-cbind(col1b,c12)
rowp<-length(Pspace3[,1])
v1<-c(rep(1,rowp*rowp))
col1<-kronecker(Pspace2,v1)
rowb<-length(Pspace2[,1])
v2<-c(rep(1,rowp))
v3<-c(rep(1,rowb))
col2<-kronecker(v3,kronecker(Pspace3,v2))
col3<-kronecker(v3,kronecker(v2,Pspace3))
SSP4<-matrix(cbind(col2,col1,col3),ncol=8)
end<-length(SSOR[,1])
g<-c(0,(p1*w21),(p1*w31),0,(p2*w12),(p2*w32),0,(p3*w13),(p3*w23))
BC<-SSP4%*%g
BC<-round(BC,5)
```

```
SSPtemp<-cbind(SSP4,BC)
SSPn<-SSP4[which(SSPtemp[,9]>=(BC0)),]
BC<-SSPn%*%g
BC<-round(BC,5)
f1<-function(p){
factorial(n1)*factorial(n2)*factorial(n3)*sum(((p[1]^(SSOR[1:(UBound),1]))/
    factorial(SSOR[1:(UBound),1]))*((p[2]^(SSOR[1:(UBound),2]))/factorial(SSOR
    [1:(UBound),2]))*
((p[3]^(SSOR[1:(UBound),3]))/factorial(SSOR[1:(UBound),3]))*((p[4]^(SSOR[1:(
    UBound),4]))/factorial(SSOR[1:(UBound),4]))*
((p[5]^(SSOR[1:(UBound),5]))/factorial(SSOR[1:(UBound),5]))*((p[6]^(SSOR[1:(
    UBound),6]))/factorial(SSOR[1:(UBound),6]))*
((p[7]^(SSOR[1:(UBound),7]))/factorial(SSOR[1:(UBound),7]))*((p[8]^(SSOR[1:(
    UBound),8]))/factorial(SSOR[1:(UBound),8])))
}
BCmatch<-which(SSOR[,9]==BChat[1])
UBound<-max(BCmatch)
BCOUT1<-apply(SSPn, 1, FUN = f1)
eval1<-BCOUT1
BCOUT<-cbind(eval1,BC)
#Test p-value - to compare to alpha
p<-max(BCOUT[which(BCOUT[,2]>=BC0),1])
```

C.1.7.2 One-Sided Test on the Difference of Two Independent BC Values, Equal Weights Only.

```
#Inputs
n1a<-#Sample Size Class 1 - CS A
n2a<-#Sample Size Class 2 - CS A
n3a<-#Sample Size Class 3 - CS A
n1b<-#Sample Size Class 1 - CS B
n2b<-#Sample Size Class 2 - CS B
n3b<-#Sample Size Class 3 - CS B
Etahat<-#Estimated BC
TV<-#Set Eta_not
##Do Not Change
##Create Sample Space
row=(n1a+1)*2
ss=matrix(seq(from=0,to=row-1,by=1), ncol=2)
ss[,2]=n1a-ss[,1]
ss1<-ss[,1:2]
row=(n2a+1)*2
ss=matrix(seq(from=0,to=row-1,by=1), ncol=2)
ss[,2]=n2a-ss[,1]
ss2<-ss[,1:2]
row=(n3a+1)*2
ss=matrix(seq(from=0,to=row-1,by=1), ncol=2)
ss[,2]=n3a-ss[,1]
ss3<-ss[,1:2]
LEN<-length(ss1[,1])*length(ss2[,1])*length(ss3[,1])
len1<-length(ss1[,1])
len2<-length(ss2[,1])
len3<-length(ss3[,1])
v1<-c(rep(1,len2*len3))
col1<-kronecker(ss1,v1)
v2<-c(rep(1,len1))
v3<-c(rep(1,len3))
col2<-kronecker(v2,kronecker(ss2,v3))
```

```
v4<-c(rep(1,len1*len2))
col3<-kronecker(v4,ss3)
SS1<-matrix(cbind(col1,col2,col3),ncol=6)
row=(n1b+1)*2
ss=matrix(seq(from=0,to=row-1,by=1), ncol=2)
ss[,2]=n1b-ss[,1]
ss1<-ss[,1:2]
row=(n2b+1)*2    ##Counting how many ways to order??
ss=matrix(seq(from=0,to=row-1,by=1), ncol=2)
ss[,2]=n2b-ss[,1]
ss2<-ss[,1:2]
row=(n3b+1)*2    ##Counting how many ways to order??
ss=matrix(seq(from=0,to=row-1,by=1), ncol=2)
ss[,2]=n3b-ss[,1]
ss3<-ss[,1:2]
LEN<-length(ss1[,1])*length(ss2[,1])*length(ss3[,1])
len1<-length(ss1[,1])
len2<-length(ss2[,1])
len3<-length(ss3[,1])
v1<-c(rep(1,len2*len3))
col1<-kronecker(ss1,v1)
v2<-c(rep(1,len1))
v3<-c(rep(1,len3))
col2<-kronecker(v2,kronecker(ss2,v3))
v4<-c(rep(1,len1*len2))
col3<-kronecker(v4,ss3)
SS2<-matrix(cbind(col1,col2,col3),ncol=6)
##Make Joint Space####
len1<-length(SS1[,1])
len2<-length(SS2[,1])
LEN<-len1*len2
v1<-c(rep(1,len2))
col1<-kronecker(SS1,v1)
v2<-c(rep(1,len1))
col2<-kronecker(v2,SS2)
SS<-matrix(cbind(col1,col2),ncol=12)
Umat<-c(1/n1a,0,1/n2a,0,1/n3a,0,-1/n1b,0,-1/n2b,0,-1/n3b,0)
U1<-SS[,1:12]%*%Umat
U1<-round(U1,5)
SS<-cbind(SS,U1)
##Order Sample Space
temp<-data.frame(SS)
SSOR<-temp[order(temp[,13]),]
#Create Prob Space to Search
pvec<-seq(from=0, to=1, by=.05)
len<-length(pvec)
v1<-c(rep(1,len*len))
col1<-kronecker(pvec,v1)
v2<-c(rep(1,len))
col2<-kronecker(v2,kronecker(pvec,v2))
col3<-kronecker(v2,kronecker(v2,pvec))
c12<-c(1-col1)
c22<-c(1-col2)
c32<-c(1-col3)
SSP1a<-cbind(col1,c12,col2,c22,col3,c32)
SSP1b<-SSP1a
len1<-length(SSP1a[,1])
```

```
len2<-length(SSP1b[,1])
LEN<-len1*len2
v1<-c(rep(1,len2))
col1<-kronecker(SSP1a,v1)
v2<-c(rep(1,len1))
col2<-kronecker(v2,SSP1b)
SSP4<-matrix(cbind(col1,col2),ncol=12)
g<-c(1,0,1,0,1,0,-1,0,-1,0,-1,0)
BC<-SSP4%*%g
BC<-round(BC,5)
SSPtemp<-cbind(SSP4,BC)
SSPna<-SSP4[which(SSPtemp[,13]==TV),]
SSPn<-SSPna
BC<-SSPn%*%g
BC<-round(BC,5)
f1<-function(p){
factorial(n1a)*factorial(n2a)*factorial(n3a)*sum(((p[1]^(SSOR[1:(LBound-1),1]))/
    factorial(SSOR[1:(LBound-1),1]))*((p[2]^(SSOR[1:(LBound-1),2]))/factorial(
    SSOR[1:(LBound-1),2]))*
((p[3]^(SSOR[1:(LBound-1),3]))/factorial(SSOR[1:(LBound-1),3]))*((p[4]^(SSOR[1:(
    LBound-1),4]))/factorial(SSOR[1:(LBound-1),4]))*
((p[5]^(SSOR[1:(LBound-1),5]))/factorial(SSOR[1:(LBound-1),5]))*((p[6]^(SSOR[1:(
    LBound-1),6]))/factorial(SSOR[1:(LBound-1),6])))
factorial(n1b)*factorial(n2b)*factorial(n3b)*sum(((p[7]^(SSOR[1:(LBound-1),7]))/
    factorial(SSOR[1:(LBound-1),7]))*((p[8]^(SSOR[1:(LBound-1),8]))/factorial(
    SSOR[1:(LBound-1),8]))*
((p[9]^(SSOR[1:(LBound-1),9]))/factorial(SSOR[1:(LBound-1),9]))*((p[10]^(SSOR
    [1:(LBound-1),10]))/factorial(SSOR[1:(LBound-1),10]))*
((p[11]^(SSOR[1:(LBound-1),11]))/factorial(SSOR[1:(LBound-1),11]))*((p[12]^(SSOR
    [1:(LBound-1),12]))/factorial(SSOR[1:(LBound-1),12])))
}
Etamatch<-which(SSOR[,13]==Etahat[1])
LBound<-min(Etamatch)
BCOUT1<-apply(SSPn, 1, FUN = f1)
eval1<-1-BCOUT1
BCOUT<-cbind(eval1,BC)
#Test p-value - to compare to alpha
p<-max(BCOUT[which(BCOUT[,2]>=TV),1])
```

Bibliography

[1] Adams, N. M. and D. J. Hand. "Comparing classifiers when the misallocation costs are uncertain". *Pattern Recognition*, 32(7):1139–1147, 1999.

[2] Agresti, A. and B. Coull. "Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions". *The American Statistician*, 52(2):119–126, May 1998.

[3] Agresti, A. *Categorical Data Analysis*. John Wiley and Sons, Inc., Hoboken, New Jersey, 2nd edition, 2002.

[4] Bache, K. and M. Lichman. "UCI Machine Learning Repository", 2013. URL http://archive.ics.uci.edu/ml.

[5] Baig, M. M., H. Gholamhosseini, and M. J. Harrison. "Fuzzy logic based anesthesia monitoring systems for the detection of absolute hypovolaemia". *Computers in Biology and Medicine*, 43:683–692, 2013.

[6] Batterton, K. A. and C. M. Schubert. "Confidence intervals around Bayes Cost in multi-state diagnostic settings to estimate optimal performance". *Statistics in Medicine*, Published Online, 2014.

[7] Blyth, C. R. and H. A. Still. "Binomial Confidence Intervals". *Journal of the American Statistical Association*, 78(381):108–116, March 1983.

[8] Boos, D. D. and L. A. Stefanski. *Essential Statistical Inference: Theory and Methods*. Springer, New York, 2013.

[9] Cantor, S. B., C. C. Sun, G. Tortolero-Luna, R. Richards-Kortum, and M. Follen. "A Comparison of C/B Ratios from Studies Using Receiver Operating Characteristic Curve Analysis". *Journal of Clinical Epidemiology*, 52(9):885–892, 1999.

[10] Canty, A. and B. Ripley. "boot: Bootstrap R (S-Plus) Functions. R package version 1.3-9.", 2013.

[11] Carpenter, J. and J. Bithell. "Bootstrap confidence intervals; when, which, what? A practical guide for medical statisticians". *Statistics in Medicine*, 19(9):1141–1164, 2000.

[12] Casella, G. and R. L. Berger. *Statistical Inference*. Duxbury, 2nd edition, 2002.

[13] Clopper, C. J. and E. S. Pearson. "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial". *Biometrika*, 26(4):404–413, December 1934.

[14] Davison, A. C. and D. Kuonen. "An Introduction to the bootstrap with applications in R". *Computing & Statistical Graphics Newsletter*, 13(1):6–11, 2002.

[15] Davison, A. C. and D. V. Hinkley. *Bootstrap Methods and their Application*. Cambridge University Press, 1997.

[16] Deng, K., C. Bourke, S. Stephen, and N. V. Vinodchandran. "New Algorithms for Optimizing Multi-Class Classifiers via ROC Surfaces". *Proceedings of the ICML 2006 Workshop on ROC Analysis in Machine Learning*. 2006.

[17] Dreiseitl, S., L. Ohno-Machado, and M. Binder. "Comparing Three-class Diagnostic Tests by Three-way ROC Analysis". *Medical Decision Making*, 20:323–331, 2000.

[18] Efron, B. "R. A. Fisher in the 21st Century". *Statistical Science*, 13(2):95–122, 1998.

[19] Fawcett, T. "An Introduction to ROC Analysis". *Pattern Recognition Letters*, 27(8):861–874, 2006.

[20] Ferri, C., J. Hernandez-Orallo, and M. A. Salido. "Volume Under the ROC Surface for Multi-class Problems". *Lecture Notes in Computer Science*, 2837:108–120, 2003.

[21] Fisher, R. A. "Inverse Probability". *Proceedings of the Cambridge Philosophical Society*, 26:528–535, July 1930.

[22] Fitxpatrick, S. and A. Scott. "Quick Simultaneous Confidence Intervals for Multinomial Proportions". *Journal of the American Statistical Association*, 82(399):875–878, September 1987.

[23] Fluss, R., D. Faraggi, and B. Reiser. "Estimation of the Youden Index and its Associated Cutoff Point". *Biometrical Journal*, 47(4):458–472, 2005.

[24] Fox, J. and S. Weisberg. *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, second edition, 2011. URL http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

[25] Gilat, A. and V. Subramaniam. *Numerical Methods for Engineers and Scientists: an Introduction with Applications using Matlab*. John Wiley and Sons, Inc., 2nd edition, 2011.

[26] Gold, R. Z. "Tests Auxiliary to χ^2 Tests in a Markov Chain". *The Annals of Mathematical Statistics*, 34(1):56–74, 1963.

[27] Goodman, L. A. "Simultaneous Confidence Intervals for Contrasts Among Multinomial Populations". *The Annals of Mathematical Statistics*, 35(2):716–725, 1964.

[28] He, X. and E. C. Frey. "The Meaning and Use of the Volume Under a Three-Class ROC Surface (VUS)". *IEEE Transactions on Medical Imaging*, 27(5):577–588, 2008.

[29] He, X., C. E. Metz, B. M. W. Tsui, J. M. Links, and E. C. Frey. "Three-Class ROC Analysis - A Decision Theoretic Approach Under the Ideal Observer Framework". *IEEE Transactions on Medical Imaging*, 25(5):571–581, 2006.

[30] Jund, J., M. Rabilloud, M. Wallon, and R. Ecochard. "Methods to Estimate the Optimal Threshold for Normally or Log-Normally Distributed Biological Tests". *Medical Decision Making*, 25:406–415, 2005.

[31] Krishnamoorthy, K. and M. Lee. "Inference for functions of parameters in discrete distributions based on fiducial approach: Binomial and Poisson cases". *Journal of Statistical Planning and Inference*, 140:1182–1192, 2010.

[32] Kvam, P. H. and B. Vidakovic. *Nonparametric Statistics with Applications to Science and Engineering*. John Wiley and Sons, Inc., Hoboken, New Jersey, 2007.

[33] Lai, C. , L. Tian, and E. F. Schisterman. "Exact Confidence Interval Estimation for the Youden Index and its Corresponding Optimal Cut-point". *Computational Statistics and Data Analysis*, 56(5):1103–1114, 2012.

[34] Lehmann, E. L. *Testing Statistical Hypotheses*. John Wiley and Sons, Inc., New York, 1959.

[35] Lo, B. W. Y., R. L. Macdonald, A. Baker, and M. A. H. Levine. "Clinical Outcome Prediction in Aneurysmal Subarachnoid Hemorrhage Using Bayesian Neural Networks with Fuzzy Logic Inferences". *Computational and Mathematical Methods in Medicine*, 2013:1–10, 2013.

[36] Luo, J. and C. Xiong. "Youden Index and Associated Cut-Points for Three Ordinal Diagnostic Groups". *Communications in Statistics-Simulation and Computation*, 42(6):1213–1234, 2013.

[37] Mas, V. R., L. A. Mas, K. J. Archer, K. Yanek, A. L. King, E. M. Gibney, A. Cotterell, R. A. Fisher, M. Posner, and D. G. Maluf. "Evaluation of Gene Panel mRNAs in Urine Samples of Kidney Transplant Recipients as a Non-invasive Tool of Graft Function". *Molecular Medicine*, 13:315–324, 2007.

[38] Mattei, T. A. "The Fuzzy Logic of Degenerative Disc Disease: From a Lorenz Attractor to a Percolation Threshold Model". *World Neurosurgery News*, 80(1-2):8–12, July/August 2013.

[39] May, W. L. and W. D. Johnson. "Constructing two-sided simultaneous confidence intervals for multinomial proportions for small counts in a large number of cells". *Journal of Statistical Software*, 5(6):1–24, 5 2000. ISSN 1548-7660. URL http://www.jstatsoft.org/v05/i06.

[40] McClish, D. K. "Evaluation of the Accuracy of Medical Tests in a Region around the Optimal Point". *Academic Radiology*, 19:1484–1490, 2012.

[41] Meeker, W. Q. and L. A. Escobar. *Statistical Methods for Reliability Data*. John Wiley and Sons, Inc., Hoboken, New Jersey, 1998.

[42] Metz, C. E. "Basic Principles of ROC Analysis". *Seminars in Nuclear Medicine*, 8(4):283–298, 1978.

[43] Molanes-Lopez, E. M. and E. Leton. "Inference of the Youden index and associated threshold using empirical likelihood for quantiles". *Statistics in Medicine*, 30:2467–2480, 2011.

[44] Mood, A. M., F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics*. McGraw-Hill, Inc., New York, 3rd edition, 1974.

[45] Nakas, C. T., T. A. Alonzo, and C. T. Yiannoutsos. "Accuracy and Cut-off Point Selection in Three-class Classification Problems Using a Generalization of the Youden Index". *Statistics in Medicine*, 29(28):2946–2955, 2010.

[46] Nakas, C. T., J. C. Dalrymple-Alford, T. J. Anderson, and T. A. Alonzo. "Generalization of Youden index for multiple-class classification problems applied to the assessment of externally validated cognition in Parkinson disease screening". *Statistics in Medicine*, 2012.

[47] Nakas, C. T. and C. T. Yiannoutsos. "Ordered multiple-class ROC analysis with continuous measurements". *Statistics in Medicine*, 23:3437–3449, 2004.

[48] Paul, L. C. "Chronic allograft nephropathy: An update". *Kidney International*, 56(3):783–793, 1999.

[49] Perkins, N. J. and E. F. Schisterman. "The Youden Index and the Optimal Cut-point Corrected for Measurement Error". *Biometrical Journal*, 47(4):428–441, 2005.

[50] Perkins, N. J. and E. F. Schisterman. "The Inconsistency of "Optimal" Cutpoints Obtained using Two Criteria based on the Receiver Operating Characteristic Curve". *American Journal of Epidemiology*, 163(7):670–675, 2006.

[51] Quesenberry, C. P. and D. C. Hurst. "Large-sample simultaneous confidence intervals for multinomial proportions". *Technometrics*, 6(2):191–195, 1964.

[52] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL http://www.R-project.org/. ISBN 3-900051-07-0.

[53] Rupert, J. and G. Miller. *Simultaneous Statistical Inference*. Springer-Verlag New York Inc., USA, second edition, 1981.

[54] Schisterman, E. F., N. J. Perkins, A. Liu, and H. Bondell. "Optimal Cut-point and its Corresponding Youden Index to Discriminate Individuals Using Pooled Blood Samples". *Epidemiology*, 16(1):73–81, 2005.

[55] Schisterman, E. F., D. Faraggi, B. Reiser, and J. Hu. "Youden Index and the Optimal Threshold for Markers with Mass at Zero". *Statistics in Medicine*, 27(2):297–315, 2008.

[56] Schisterman, E. F. and N. Perkins. "Confidence Intervals for the Youden Index and Corresponding Optimal Cut-Point". *Communications in Statistics: Simulation and Computation*, 36(3):549–563, 2007.

[57] Schubert, C. M. and T. Guennel. "Comparing performance of multi-class classification systems with ROC manifolds: When volume and correct classification fails". *Communications in Statistics-Simulation and Computation:ACCEPTED*.

[58] Schubert, C. M., S. N. Thorsen, and M. E. Oxley. "The ROC Manifold for Classification Systems". *Pattern Recognition*, 44(2):350–362, 2011.

[59] Selvan, S., M. Kavitha, S. Shenbaga Devi, and S. Suresh. "Fuzzy-Based Classification of Breast Lesions Using Ultrasound Echography and Elastography". *Ultrasound Quarterly*, 28(3):159–167, September 2012.

[60] Sen, A. and M. Srivastava. *Regression Analysis: Theory, Methods, and Applications*. Springer-Verlang New York Inc., New York, NY, 1990.

[61] da Silva, J. Estrela, J. P Marques de Sa, and J. Jossinet. "Classification of Breast Tissue by Electrical Impedance Spectroscopy". *Medical and Biological Engineering and Computing*, 38:26–30, 2000.

[62] Sison, C. P. and J. Glaz. "Simultaneous Confidence Intervals and Sample Size Determination for Multinomial Proportions". *Journal of the American Statistical Association*, 90(429):366–369, March 1995.

[63] Skaltsa, K. , L. Jover, and J. L. Carrasco. "Diagnostic Tests Optimum Threshold: Accounting for Decision Costs and Uncertainty Around the Cut-Off". *Proceedings of 22th Panhellenic Conference on Statistics*, 333–340. 2009.

[64] Skaltsa, K. , L. Jover, and J. L. Carrasco. "Estimation of the Diagnostic Threshold Accounting for Decision Costs and Sampling Uncertainty". *Biometrical Journal*, 52(5), 2010.

[65] Skaltsa, K. , L. Jover, D. Fuster, and J. L. Carrasco. "Optimum Threshold Estimation Based on Cost Function in a Multistate Diagnostic Setting". *Statistics in Medicine*, 31:1098–1109, 2012.

[66] Subtil, F. "Comments on the article of C.Y.Lai, L.Tain, and E.F.Schisterman on the "Exact confidence interval estimation for the Youden index and its corresponding optimal cut-point"". *Computational Statistics and Data Analysis*, 55:3379–3380, 2011.

[67] Sunshine, J. "Contributed Comment". *Academic Radiology*, 2:S72–S74, 1995.

[68] Swets, J. A., R. M. Dawes, and J. Monahan. "Better Decisions through Science". *Scientific American*, 82–87, 2000.

[69] Villacorta, P. J. *MultinomialCI: Simultaneous confidence intervals for multinomial proportions according to the method by Sison and Glaz*, 2012. URL http://CRAN.R-project.org/package=MultinomialCI. R package version 1.0.

[70] Wald, A. "Tests of statistical hypotheses concerning several parameters when the number of observations is large". *Transactions of the American Mathematical Society*, 54(3):426–482, November 1943.

[71] Wang, H. "Exact confidence coefficients of simultaneous confidence intervals for multinomial proportions". *Journal of Multivariate Analysis*, 99(5):896–911, 2008.

[72] Wang, Y. H. "Fiducial Intervals: What Are They?" *The American Statistician*, 54(2):105–111, May 2000.

[73] Weerahandi, S. *Exact Statistical Methods for Data Analysis*. Springer-Verlag New York, Inc., New York, 1994.

[74] Xinmin, L., L. Guoying, and X. Xingzhong. "Fiducial intervals of restricted parameters and their applications". *Science in China Ser. A Mathematics*, 48(11):1567–1583, 2005.

[75] Xiong, C., G. van Belle, J. P. Miller, and J. C. Morris. "Measuring and Estimating Diagnostic Accuracy When There Are Three Ordinal Diagnostic Groups". *Statistics in Medicine*, 25(7):1251–1273, 2006.

[76] Youden, W. J. "Index for Rating Diagnostic Tests". *Cancer*, 3(1):32–35, 1950.

[77] Zabell, S. L. "R. A. Fisher and the Fiducial Argument". *Statistical Science*, 7(3):369–387, 1992.

[78] Zhao, S., X. Xu, and X. Ding. "Fiducial inference under nonparametric situations". *Journal of Statistical Planning and Inference*, 142:2779–2798, 2012.

[79] Zhou, H. and G. Qin. "New Nonparametric Confidence Intervals for the Youden Index". *Journal of Biopharmaceutical Statistics*, 22(6):1244–1257, 2012.

[80] Zhou, H. and G. Qin. "Confidence intervals for the difference in paired Youden indices". *Pharmaceutical Statistics*, 12:17–27, 2013.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
18–09–2014	Doctoral Dissertation	Sept 2011–Sept 2014

4. TITLE AND SUBTITLE

Statistical Inference on Optimal Points to Evaluate Multi-State Classification Systems

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Batterton, Katherine Anne, Captain, USAF

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–ENC–DS–14–S–02

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
Distribution Statement A: Approved for public release; Distribution unlimited

13. SUPPLEMENTARY NOTES
This work is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

14. ABSTRACT
In decision making, an optimal point represents the settings for which a classification system should be operated to achieve maximum performance. Clearly, these optimal points are of great importance in classification theory. Not only is the selection of the optimal point of interest, but quantifying the uncertainty in the optimal point and its performance is also important.

The Youden index is a metric currently employed for selection and performance quantification of optimal points for classification system families. The Youden index quantifies the correct classification rates of a classification system, and its confidence interval quantifies the uncertainty in this measurement. This metric currently focuses on two or three classes, and only allows for the utility of correct classifications and the cost of total misclassifications to be considered. An alternative to this metric for three or more classes is a cost function which considers the sum of incorrect classification rates. This new metric is preferable as it can include class prevalences and costs associated with every classification. In multi-class settings this informs better decisions and inferences on optimal points.

The work in this dissertation develops theory and methods for confidence intervals on a metric based on misclassification rates, Bayes Cost, and where possible, the thresholds found for an optimal point using Bayes Cost. Hypothesis tests for Bayes Cost are also developed to test a classification systems performance or compare systems with an emphasis on classification systems involving three or more classes. Performance of the newly proposed methods is demonstrated with simulation.

15. SUBJECT TERMS
optimal point, ROC, confidence intervals, classification system

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr.Christine Schubert Kabban (ENC)
U	U	U	UU	219	19b. TELEPHONE NUMBER *(include area code)* (937) 255-3636x4549 christine.schubertkabban@afit.edu