Dissertations & Theses (Open Access)   School of Public Health

5-2020

Statistical Evaluation Of An Extension Of Linear Discriminant Analysis In Multiclass Settings With An Application For Detection Of Cervical Neoplasia

Frances A. Brito
UTHealth School of Public Health

STATISTICAL EVALUATION OF AN EXTENSION OF LINEAR DISCRIMINANT

ANALYSIS IN MULTICLASS SETTINGS WITH AN APPLICATION FOR DETECTION

OF CERVICAL NEOPLASIA

by

FRANCES A BRITO, BS

APPROVED:

JOSE-MIGUEL YAMAL, MA, PHD

JOSE-MIGUEL YAMAL, MA, PHD

CRAIG HANIS, MA, PHD

STATISTICAL EVALUATION OF AN EXTENSION OF LINEAR DISCRIMINANT

ANALYSIS IN MULTICLASS SETTINGS WITH AN APPLICATION FOR DETECTION

OF CERVICAL NEOPLASIA

by

FRANCES A BRITO
BS, University of Georgia, 2012

Presented to the Faculty of The University of Texas

School of Public Health

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH
Houston, Texas
May, 2020

ACKNOWLEDGEMENTS

I wish to express my deepest gratitude to my thesis chair and academic advisor, Dr. Jose-Miguel Yamal. His guidance and instruction undoubtedly prepared me to be a proficient, diligent, and well-rounded researcher. Completion of this thesis would not be possible without his constant support and knowledge. I would also like to recognize, Dr. Craig Hanis, who not only offered valuable comments and insights for my thesis, but also advice and encouragement for future career goals.

I would like to thank my parents, whose love and guidance are with me in whatever I pursue. I am forever thankful for my boyfriend, Thomas, who inspired me to pursue a graduate degree and whose unwavering support and advice was invaluable. Special thanks to both my friends within the UTHealth community, whose sense of comradery was unparalleled, and lifelong friends who provided pleasant distractions and empathy throughout this journey.

STATISTICAL EVALUATION OF AN EXTENSION OF LINEAR DISCRIMINANT

ANALYSIS IN MULTICLASS SETTINGS WITH AN APPLICATION FOR DETECTION

OF CERVICAL NEOPLASIA

Frances A Brito, BS, MS
The University of Texas
School of Public Health, 2020

Thesis Chair: Jose-Miguel Yamal, MA, PhD

Although a pathologists' review of Papanicolaou smear cell samples has been

successful in decreasing cervical cancer incidence, it is often costly and time-consuming.

Quantitative cytology is a promising semi-automated method that measures cell features for

further analysis or classification. There have been several advancements in classification

algorithms, but many do not account for the nested data structure seen in quantitative

cytology. Further, histologic diagnoses are separated into five or more classes, yet, multi-

class classification has not been investigated. Here, we compare the predictive performance

of macrolevel discriminant analysis (MDA) to traditional discriminant analysis methods in

multi-class settings on cervical quantitative cytology data and simulated data sets. MDA had

similar overall classification accuracy and area under the ROC curve results to linear

discriminant analysis (LDA) and quadratic discriminant analysis (QDA) when applied to

cervical quantitative cytology data. However, MDA has a tremendous advantage over LDA

and QDA methods when a macrolevel (patient or individual) effect is assumed and when one

class is composed of a mixture of gaussian distributions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

BACKGROUND

Introduction

Approximately 583,000 cervical cancer cases are discovered each year, accounting for 7.9% of all female cancer incidences. Worldwide, 266,000 annual deaths are due to cervical cancer [1]. Treatment of cervical cancer is highly successful with an expected 91% 5-year survival rate if detected while the cancer is localized and greater than 91% if detected in the pre-cancerous stage [2]. The current standard for detection of premalignant lesions is through the use of Papanicolaou smears in which a sample of cells is obtained from a patient's cervix. A pathologist then examines these cells under a microscope and classifies a patient into one of several possible varying stages of disease including, but not limited to: cancer, high-grade, low-grade, atypical squamous cells of undetermined significance (ASCUS), or negative for dysplasia. One promising alternative to the costly and onerous Papanicolaou smear screening process is through the use of quantitative cytology. This semi-automated technology classifies cells by measuring features on cell images. There have been several advancements in classification of the stages of cervical cancer due to computer automation and machine learning algorithms. However, many procedures only have the capability of predicting up to 2 classes. Because the histologic diagnosis is actually ordinal, it is usually dichotomized to be used in those methods.

Many attempts have been made to apply classical discriminant analysis to repeated measures data, but fall short of accurate classification due to several factors: missing values of a single observation results in deletion of entire individuals, application to high dimensional data (i.e. $\# \ of \ individuals \ll p \times \# \ of \ total \ observations$) is difficult if not

impossible, and central measures of tendency of the individual are compared to class means and variances with no regard to variation of observations within an individual [3]. Yamal et al. (currently under review) provides a solution to classifying data with a nested structure where multiple measurements are obtained per subject and correlations exist among measurements within a subject. This method is outlined here.

First, assume that we wish to classify an observation into one of two classes (class r or l) using p predictors. Let X be a random vector of predictors or features and Y be the random response variable that can take on the value r or l. Additionally, let π_r be the prior probability that a given observation is associated with the rth class. Applying Bayes' theorem, we obtain

$$\Pr(Y = r | X = x) = \frac{\pi_r f_r(x)}{\pi_r f_r(x) + \pi_l f_l(x)} \tag{1-1}$$

where $f_r(x)$ is the class conditional density of X. Linear discriminant analysis (LDA) makes use of the assumption that observations in the kth class are drawn from a multivariate normal gaussian distribution where $X \sim N(\mu_k, \Sigma)$ such that

$$f_k(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x - \mu_k)^T \Sigma^{-1}(x - \mu_k)} \tag{1-2}$$

Furthermore, LDA assumes that Σ is a covariance matrix that is common to all K classes. Thus,

$$\Pr(Y = r | X = x) = \frac{\pi_r \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x - \mu_r)^T \Sigma^{-1}(x - \mu_r)}}{\pi_r \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x - \mu_r)^T \Sigma^{-1}(x - \mu_r)} + \pi_l \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x - \mu_l)^T \Sigma^{-1}(x - \mu_l)}} \tag{1-3}$$

simplifies to

$$\frac{\pi_r e^{-\frac{1}{2}(x-\mu_r)^T \Sigma^{-1}(x-\mu_r)}}{\pi_r e^{-\frac{1}{2}(x-\mu_r)^T \Sigma^{-1}(x-\mu_r)} + \pi_l e^{-\frac{1}{2}(x-\mu_l)^T \Sigma^{-1}(x-\mu_l)}} \tag{1-4}$$

Taking the log of this simplified equation provides us the discriminant function where assignment of an observation, $X = x$, is to the class for which

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + log \pi_k \tag{1-5}$$

is largest for class r or l.

To demonstrate how Yamal et al. utilized the well-known classification methods of LDA, an implementation of macrolevel discriminant analysis (MDA) will be provided in the context of quantitative cytology data. Assume that $p=2$ features are measured on each cell and the patient (referring to the population of cells for a patient) will be classified into one of two k classes, r or l. Here, we may denote patients as the macrolevel and cells as the microlevel.

Let X_{ij} be a vector of random variables for p features of the selected cell i of patient j. X_{ij} can be constructed out of the sum of two distributions Z_{ij} and U_j. The distribution of all cells in the population given a class k is $f(Z_{ij}|Y_j = k) \sim MVN(\mu_k, V_z)$ where the covariance V_Z is a $p \times p$ matrix. U_j is the patient effect (a vector of length p) measuring the deviation of the patient feature means from the population feature means and is distributed as a $MVN(0, V_U)$ where the covariance V_U is a $p \times p$ matrix. Now, let X_j be a vector of length $p \times n_j$ where the features of each cell for a patient are vectorized and concatenated, and n_j is the number of cells for patient j. Thus, the conditional distribution is $f(X_j|Y_j =$

3

$k) \sim MVN(\boldsymbol{\mu_k}, V_X^j)$ where $\boldsymbol{\mu_k}$ is a column vector of length $p \times n_j$ composed of the mean of

each feature for class k and repeated n_j times. Lastly:

$$V_X^j = \begin{pmatrix} V_Z + V_U & V_U & \cdots & V_U \\ V_U & V_Z + V_U & \cdots & V_U \\ \vdots & \vdots & \ddots & \vdots \\ V_U & V_U & \cdots & V_Z + V_U \end{pmatrix} \tag{1-6}$$

This covariance matrix is not the same for each patient due to the varying random number of

cells obtained for each patient, but, as with traditional LDA, we can assume that each cell has

a common covariance. Thus, using the LDA function from above:

$$\Pr(Y_j = r | X_j = x_j) = \frac{\pi_r e^{-\frac{1}{2}(x_j - \mu_r)^T (V_X^j)^{-1}(x_j - \mu_r)}}{\pi_r e^{-\frac{1}{2}(x_j - \mu_r)^T (V_X^j)^{-1}(x_j - \mu_r)} + \pi_l e^{-\frac{1}{2}(x_j - \mu_l)^T (V_X^j)^{-1}(x_j - \mu_l)}} \tag{1-7}$$

and similarly, we classify a patient, $X_j = x_j$, to the class for which the discriminant function

$$\delta_k(x_j) = x_j^T (V_X^j)^{-1} \mu_k - \frac{1}{2} \mu_k^T (V_X^j)^{-1} \mu_k + log\pi_k \tag{1-8}$$

is largest.

Parameter estimates are obtained from a sample as follows:

- $\hat{\pi}_k = \frac{n_k}{N}$ where n_k is the number of patients associated with class k and N is the total

  number of patients in the sample

- $\hat{\mu}_k = \frac{1}{n_k} \Sigma_{\{j:Y_j=k\}} x_i$ is the sample mean of all cells for patients belonging to class k

- x_j is a vector of all measured features of cells obtained from a patient and $\bar{x}_j$ is a

  vector of length p where each value is the mean of each feature of all cells for a

  patient

4

- $\widehat{U}_j = \bar{x}_j - \hat{\mu}_k$ is the estimated patient effect where the class mean for a patient belonging to class k is subtracted from the patient mean; These estimates are concatenated to form a matrix with dimension $N \times p$

- $\widehat{V}_u = \frac{1}{N-k}\sum_{k=1}^{K}\sum_{j:Y_j=k}(\bar{x}_j - \hat{\mu}_k)(\bar{x}_j - \hat{\mu}_k)^T$ is the between patient variation

- $\hat{Z}_j = x_{ij} - \bar{x}_j$ subtracts the patient mean from each cell within a patient; These estimates are concatenated to form a matrix with dimension $number\ of\ total\ cells\ in\ sample \times p$

- $\widehat{V}_Z = \frac{1}{N_{cells}}\sum_j\sum_i(x_{ij} - \bar{x}_j)(x_{ij} - \bar{x}_j)^T$ is the within patient variation

A significant advantage of LDA has been the ability to reduce a higher-dimensional problem into lower dimensions (1) to directly classify objects (as described above); (2) to aid visualization; and (3) to be used as new features by projecting a feature space into a subspace while keeping the discrimination information. Dimension reduction helps to avoid overfitting due to subsequently having fewer parameters to estimate. It also helps significantly with computational costs. The MDA method is expected to enjoy the same benefits, but these have not been explored. Specifically, the MDA method has not been evaluated via simulations for more than two classes. In previous analyses, patients were predicted into one of two classes: high grade or worse vs. low grade or better. However, there are actually 5 ordinal histologic categories in which a patients' sample of cells could be classified. This thesis evaluates and compares the classification accuracy of MDA for two and three classes in a series of simulations. We also applied this method to cervical quantitative cytology data to classify into more than two histologic diagnoses, including the use of MDA as a visualization tool.

Public Health Significance

Pathologists' reviews of Papanicolaou smears are not only costly and labor intensive, but are subjective and relies on proper education, training, and resources, many of which are not available in underdeveloped countries. Furthermore, high interpretive variability exists even among expert pathologists [4] and thus standardization of this process would be advantageous. Current classification algorithms for macrolevel data have primarily classified patients into two classes: high-grade disease or worse and low-grade disease or normal. However, women with low-grade lesions have an increased risk of developing more advanced stages of pre-cancerous lesions [5] and require, at a minimum, closer monitoring than those with normal cells. Additionally, personalized treatment based on precise diagnoses may reduce risk of complications due to overtreatment such as premature labor in the future [6].

Accurate and precise classification and prediction of disease severity is essential for determining the types, length, and intensity of treatment for patients across the healthcare industry. For example, in one study, successful risk-based classification led to decreased toxicity and improved outcome for low- and intermediate-risk patients, as well as higher survival rates for high-risk patients. There is a total of 16 risk groups for patients with neuroblastoma and classification is crucial at onset [7]. Furthermore, automated classification removes human error and ensures consistency across hospitals and even countries. On a broader scope, there has been an increasing need to classify groups, neighborhoods etc. in the public health sector, such as analysis of cluster randomized trials [8].

Specific Aims

1. Conduct computer simulations of the macrolevel discriminant analysis method predicting two and three classes under varying circumstances to assess model behavior.

2. Apply the macrolevel discriminant analysis model to quantitative cytology data to predict two, three, and four class groupings of stages of cervical cancer and assess prediction accuracy.

METHODS

Quadratic MDA

Typical LDA and linear MDA assume a common covariance matrix among the classes. However, similar to quadratic discriminant analysis, our expansion to multiple class classification required a unique covariance for each class. The simplified LDA equation (1-4) no longer applies and the posterior probability of a patient being in class k is given by:

$$\Pr\left(Y_j = k \middle| X_j = x_j\right) = \frac{\pi_k \dfrac{1}{(2\pi)^{\frac{p}{2}}\left|V_X^{j,k}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_j-\mu_k)^T\left(V_X^{j,k}\right)^{-1}(x_j-\mu_k)}}{\sum_{i=1}^{K} \pi_k \dfrac{1}{(2\pi)^{\frac{p}{2}}\left|V_X^{j,k}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_j-\mu_k)^T\left(V_X^{j,k}\right)^{-1}(x_j-\mu_k)}} \tag{2-1}$$

and the quadratic MDA function becomes

$$\delta_k(x) = -\frac{1}{2}log\left|V_X^{j,k}\right| - \frac{1}{2}(x - \mu_k)^T \ (V_X^{j,k})^{-1}(x - \mu_k) + log\pi_k \tag{2-2}$$

Where, again, a patient is classified into the class for which the quadratic MDA function is greatest. Parameter estimates are similar to those described previously but are now class specific:

- $\widehat{V_u}^k$ becomes $\widehat{V}_u^k = \frac{1}{n_k}\sum_{j:Y_j=k}\left(\bar{x}_j - \hat{\mu}_k\right)\left(\bar{x}_j - \hat{\mu}_k\right)^T$;

7

- $\hat{V}_z$ becomes $\hat{V}_z^k = \frac{1}{n_{k,cells}} \sum_{j:Y_j=k} \sum_i (x_{ij} - \bar{x}_j)(x_{ij} - \bar{x}_j)^T$; and

- V_X^j becomes

$$V_X^{j,k} = \begin{pmatrix} V_Z^k + V_U^k & V_U^k & \cdots & V_U^k \\ V_U^k & V_Z^k + V_U^k & \cdots & V_U^k \\ \vdots & \vdots & \ddots & \vdots \\ V_U^k & V_U^k & \cdots & V_Z^k + V_U^k \end{pmatrix}$$

Regularized MDA

Due to the high dimensionality and increased correlation of nested data, the estimated

covariance matrices are often unstable and singular. Therefore, we introduced a

regularization technique presented in Guo et al. [9]. Here, the within patient and between

patient covariance matrices are regularized using the identity matrix:

$$\tilde{V}_u^k = \alpha V_u^k + (1 - \alpha)I_p \qquad (2\text{-}3)$$

and

$$\tilde{V}_z^k = \alpha V_z^k + (1 - \alpha)I_p \qquad (2\text{-}4)$$

for some $\alpha, 0 \leq \alpha \leq 1$. Given that the V_Z^k and V_U^k matrices are additive, having the same

regularization parameter is a reasonable simplification to train the model. An alternative

model could use separate parameters for each matrix.

Cervical quantitative cytology data application

The data available includes quantitative measurements on an average of 2600 cells

collected from 1728 women. A patient's disease status was determined by the worst

histologic grade out of all biopsies. Further details on data collection may be found in Yamal et al. and Guillaud et al. [10,11]. Histological diagnosis classes are defined in Table 1.

Table 1: Cervical data histological diagnosis class allocations

Class	Histological diagnosis
1	Negative for dysplasia
2	Atypical squamous cells of undetermined significance (ASCUS)
3	HPV associated Cancer (HPVaC) Mild
4	Moderate Severe
5	Carcinoma In Situ (CIS) Cancer

For this analysis, we combined class 2 and 3 & class 4 and 5 when classifying into 3 categories. Class 4 and 5 were combined when classifying into 4 categories. Previous studies suggest that DNA Index and optical density (OD) skewness are associated with disease and thus will be the sole variables used in all analyses [10,11].

Variable reduction methods

Fortunately, previous studies have identified important predictor variables to be used in this analysis. However, this information is not always available to researchers and other variable selection methods need to be employed. Lasso is one method of variable reduction that places a constraint on the coefficients so that the estimates are shrunk. Some coefficient estimates are shrunk to zero thus, only variables with non-zero coefficients will be considered in the model [12].

Principal component analysis (PCA) provides an alternate option to variable selection by reducing dimensionality. Although not typically a concern in prediction models, this does

come at the expense of interpretability. PCA produces a set of new uncorrelated variables (components) each of which represent a linear combination of the original variables. The first principal component is found by accounting for the largest amount of variability in the data. The components that follow will also maximize variance but are subject to the constraint of being orthogonal (uncorrelated) with prior components [12].

There are several other methods of variable reduction that may be chosen based on the type of data or problem one would like to solve.

Assessment of classification accuracy in cervical cancer data

The cervical cancer data was randomly divided into a training set (40%), validation set (30%) to tune parameters, and a test set (30%). Three classification methods were considered: (1) Macrolevel Discriminant Analysis as described above; (2) means of each of the two cell features are computed for each patient and subsequently classified using Linear Discriminant Analysis; and (3) means of each of the two cell features are computed for each patient and subsequently classified using Quadratic Discriminant Analysis.

An average test set accuracy was computed for each classification method as follows:

$$\frac{\sum_{i=1}^{k} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i}}{k} \tag{2-5}$$

where k is the total number of classes, tp is the number of true positives, tn is the number of true negatives, fn is the number of false negatives, and fp is the number of false positives. Additionally, we calculated the area under the Receiver Operating Characteristic (ROC) curve (AUC) as another measure of classification accuracy. This measure is typically used in

two-class supervised classification settings, but was extended to be used in the multiclass

setting by averaging the AUC's of pairwise comparisons of the classes [13].

Simulations

We conducted several simulation studies and compared overall classification

accuracies and AUC's across the three methods previously described: MDA, LDA at the

patient level, and QDA at the patient level. Each simulation utilized two predictor variables

(features) under two scenarios; classification into two and three classes. Due to computation

constraints we limited the number of patients to 250 per class with 500 cells per patient.

The first simulation assumed the model such that there is a macro level effect

$U_j \sim MVN(0, V_U)$, $Z_j \sim MVN(\mu_k, V_z)$, and $X_j = U_j + Z_j$. Where mean vectors and

covariance matrices were defined as follows for each class:

Table 2: Simulation 1 parameters by class

Class 1	$\mu_1 = (0,0)$	$V_u^1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$V_z^1 = \begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix}$
Class 2	$\mu_2 = (0.05, -0.05)$	$V_u^2 = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 2 \end{bmatrix}$	$V_z^2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
Class 3 (when including a third class)	$\mu_3 = (-0.05, 0.05)$	$V_u^3 = \begin{bmatrix} 2 & -0.5 \\ -0.5 & 2 \end{bmatrix}$	$V_z^3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

Figure 1 provides a visualization of a random sample of six patients (two for each class) and

100 cells per patient. The ellipse captures 95% of the data points for each patient.
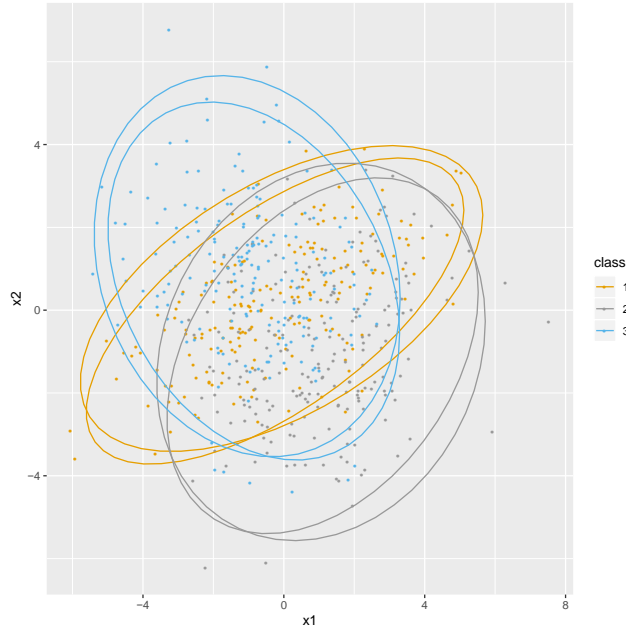
11

Figure 1: Simulation 1 sample data visualization

In the second simulation, we assumed the same model as simulation 1. However, we increased the number of macrolevel subjects incrementally from 10 to 1000 and the number of microlevel observations from 10 to 2000 in order to assess the influence of small to large microlevel and macrolevel sample sizes.

In the remaining simulations we investigated prediction accuracy across the three methods when the model is misspecified. The third simulation included a right-skewed variable in each of the classes and thus, violating the normality assumption required for LDA. Mean vectors for each class were adjusted to implement this simulation so that $\mu_1 = (5,5)$, $\mu_2 = (5.05, 4.95)$, and $x_{1j} = x_{1j}^2$. The fourth simulation contains a class that is composed of a mixture of Gaussian distributions. For the two-class scenario: if $Y_j = 1$,

$V_z \sim MVN((3,3)^T, \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix})$, $V_u \sim MVN((0,0)^T, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix})$ and if $Y_j = 2$,

$V_z \sim 0.5 MVN\left((3.5,2.5)^T, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) + 0.5(MVN\left((2.5,3.5)^T, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$,

12

$V_u \sim 0.5 MVN\left((0,0)^T, \begin{bmatrix} 3.5 & 1.5 \\ 1.5 & 2.5 \end{bmatrix}\right) + 0.5\left(MVN\left((0,0)^T, \begin{bmatrix} 3.5 & 1.5 \\ 1.5 & 2.5 \end{bmatrix}\right)\right)$ For the three-class

scenario: if $Y_j = 1$, $V_z \sim MVN((3,3)^T, \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix})$, $V_u \sim MVN((0,0)^T, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix})$; if $Y_j = 2$,

$V_z \sim 0.5 MVN\left((3.2,3)^T, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) + 0.5\left(MVN\left((3,3.2)^T, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)\right)$,

$V_u \sim 0.5 MVN\left((0,0)^T, \begin{bmatrix} 3.5 & 1.75 \\ 1.75 & 2.5 \end{bmatrix}\right) + 0.5\left(MVN\left((0,0)^T, \begin{bmatrix} 3.5 & 1.75 \\ 1.75 & 2.5 \end{bmatrix}\right)\right)$; if $Y_j = 3$,

$V_z \sim 0.5 MVN\left((3,3)^T, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) + 0.5\left(MVN\left((3.2,3.2)^T, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)\right)$,

$V_u \sim 0.5 MVN\left((0,0)^T, \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}\right) + 0.5\left(MVN\left((0,0)^T, \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}\right)\right)$. The idea is to create a

bimodal distribution in class two and three such that the combined distributions have mean

vectors and covariance matrices similar to class one. Figure 2 provides a visualization of

sample data for the two-class scenario. Here, class 1 consists of two individuals; each

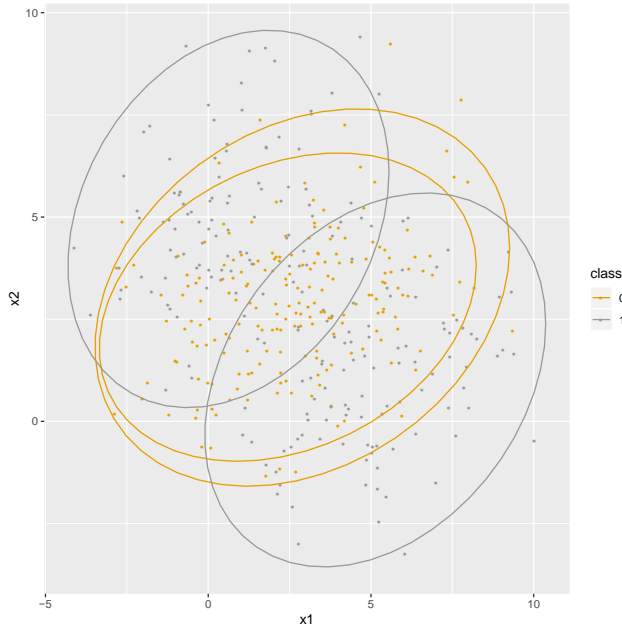sampled from two gaussian distributions and belong to the same class.

Figure 2: Simulation 4 sample data visualization

Lastly, in the fifth simulation, we utilized the same parameters shown in Table 2. However, we began with covariance matrices with zero values in the off diagonals, i.e. no covariance, and incrementally increased the covariance between the two features in all classes. Specifically, V_z^1 was initialized as $\begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$ and the off diagonals increased by 0.5 in each iteration up to $\begin{bmatrix} 4 & 2.5 \\ 2.5 & 1 \end{bmatrix}$; V_u^2 was initialized as $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$ and the off diagonals increased by 0.1 in each iteration up to $\begin{bmatrix} 2 & 0.5 \\ 0.5 & 2 \end{bmatrix}$; V_u^3 was initialized as $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$ and the off diagonals decreased by 0.1 in each iteration to $\begin{bmatrix} 2 & -0.5 \\ -0.5 & 2 \end{bmatrix}$. Identity matrices were set for all other covariance matrices (V_u^1, V_z^2, V_z^3). Figure 3 provides a visualization of sample data that has no covariation between two features transitioning to sample data that has high covariation. Here, the sample data consists of 2 patients per class with 100 cells per patient and each ellipse encompasses 95% of the data points for each class.

Figure 3: Simulation 5 sample data visualization
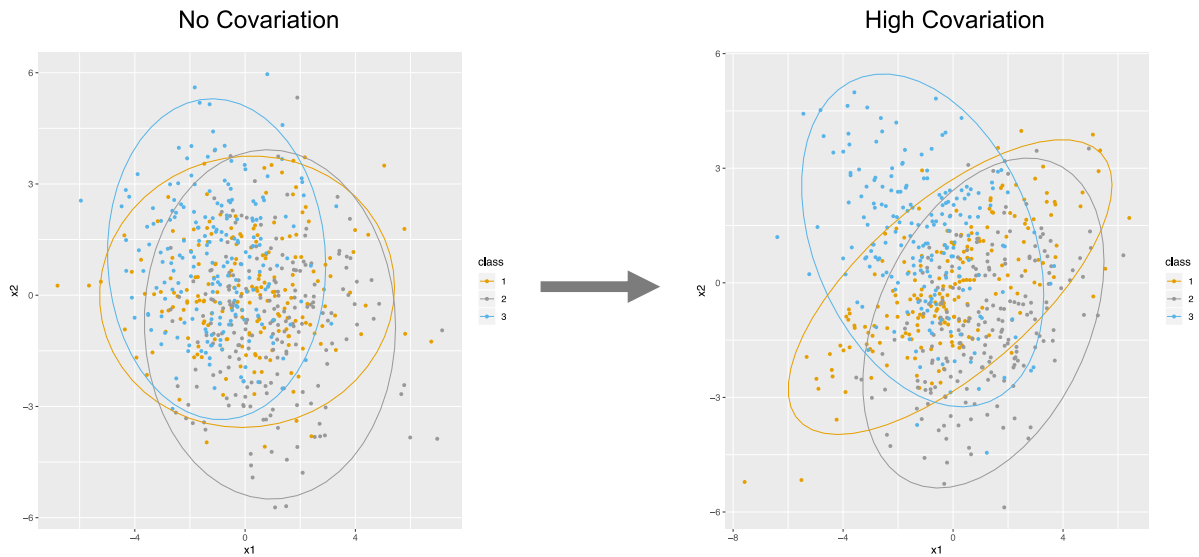
Assessment of classification accuracy in simulations

100 separate train and test sets were created for each simulation. The distributions of overall classification accuracies and AUC's were recorded. In total, we had 6 accuracy distributions for each simulation – one for each combination of the three methods used and the number of classes attempting to predict (two and three).

All analyses were performed using the statistical package R version 3.5 (R Foundation for Statistical Computing, Vienna, Austria).

IRB Approval

This research received IRB approval ("Optical Technologies and Molecular Imaging of Cervical Neoplasia", HSC-SPH-10-0631). All data was de-identified and analyzed on a secure UTHealth server.

RESULTS

Simulation Results

The distributions of overall classification accuracy and AUC after repeating simulation 1 (assumption of macrolevel effect) 100 times are displayed in figures 4 and 5. MDA had an observable advantage over LDA and QDA with nearly perfect prediction accuracy on the test sets.

In simulation 2, increasing the number of macrolevel individuals resulted in small changes to the overall classification accuracy or AUC. However, these accuracy measurements increased drastically as the number of microlevel observations increased (Tables 6 & 7). This trend was observed in both two and three-class scenarios.

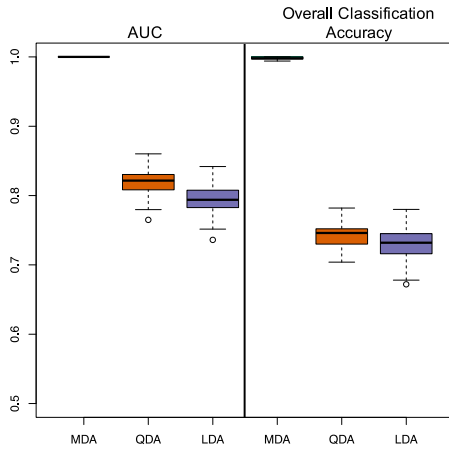Figure 4: Simulation 1 two-class test set accuracies
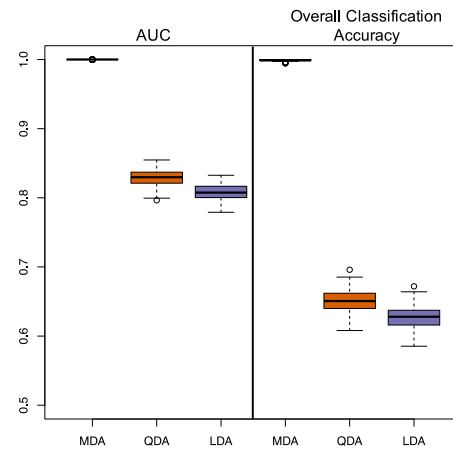

Figure 5: Simulation 1 three-class test set accuracies

Table 6: Simulation 2 two-class overall classification accuracy and area under the ROC curve using MDA on test set varying the number of macro-level and micro-level observations

		Number of Patients per Class					
		10	20	50	100	200	1000
		Overall Classification Accuracy					
	10	0.790	0.785	0.807	0.809	0.812	0.808
	20	0.818	0.838	0.839	0.849	0.853	0.851
	50	0.863	0.891	0.888	0.900	0.902	0.901
	100	0.930	0.929	0.939	0.937	0.943	0.945
	150	0.944	0.957	0.962	0.964	0.963	0.966
	250	0.975	0.977	0.984	0.985	0.986	0.987
Number of Cells per Patient	**350**	0.987	0.991	0.993	0.993	0.994	0.994
	450	0.996	0.995	0.997	0.997	0.997	0.998
	2000	1.000	1.000	1.000	1.000	1.000	1.000
		AUC					
	10	0.948	0.949	0.955	0.955	0.955	0.955
	20	0.990	0.991	0.993	0.993	0.993	0.992
	50	1.000	1.000	1.000	1.000	1.000	1.000
	100	1.000	1.000	1.000	1.000	1.000	1.000
	150	1.000	1.000	1.000	1.000	1.000	1.000
	250	1.000	1.000	1.000	1.000	1.000	1.000
	350	1.000	1.000	1.000	1.000	1.000	1.000
	450	1.000	1.000	1.000	1.000	1.000	1.000
	2000	1.000	1.000	1.000	1.000	1.000	1.000

Table 7: Simulation 2 three-class overall classification accuracy and area under the ROC curve using MDA on test set varying the number of macro-level and micro-level observations

	Number of Patients per Class					
	10	20	50	100	200	1000
	Overall Classification Accuracy					
10	0.583	0.614	0.645	0.655	0.663	0.671
20	0.653	0.687	0.709	0.740	0.741	0.749
50	0.749	0.797	0.824	0.843	0.852	0.858
100	0.839	0.894	0.911	0.917	0.929	0.933
150	0.884	0.930	0.953	0.955	0.962	0.966
250	0.926	0.974	0.983	0.986	0.988	0.988
350	0.966	0.985	0.993	0.995	0.996	0.996
450	0.984	0.994	0.998	0.998	0.998	0.998
2000	1.000	1.000	1.000	1.000	1.000	1.000
	AUC					
10	0.834	0.854	0.861	0.864	0.868	0.870
20	0.894	0.912	0.915	0.921	0.919	0.921
50	0.959	0.962	0.966	0.969	0.970	0.970
100	0.985	0.990	0.991	0.991	0.992	0.992
150	0.993	0.996	0.997	0.997	0.997	0.998
250	0.999	0.999	1.000	1.000	1.000	1.000
350	1.000	1.000	1.000	1.000	1.000	1.000
450	1.000	1.000	1.000	1.000	1.000	1.000
2000	1.000	1.000	1.000	1.000	1.000	1.000

Number of Cells per Patient (vertical axis label)

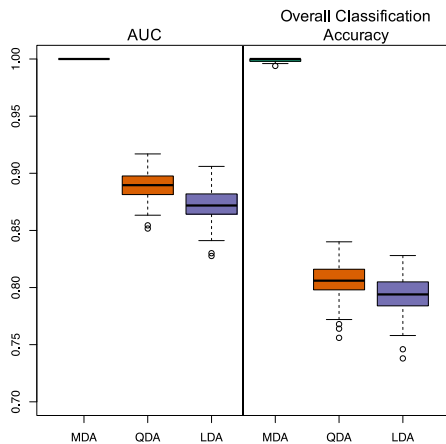Figure 6: Simulation 3 two-class test set accuracies in simulation 3
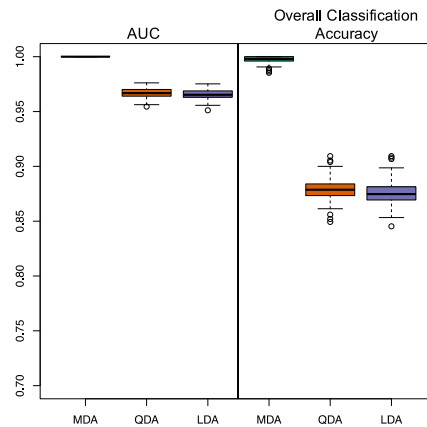
Figure 7: Simulation 3 three-class test set accuracies

17

Simulation three results were similar to those observed in simulation one. MDA had an advantage and higher overall classification accuracy and AUC measurements over LDA and QDA even when the assumption of normality was violated (Figures 6 & 7).

Introducing a class with a mixture of gaussian distributions resulted in decreased accuracies across all models. MDA had a large variation of overall classification accuracies, but an AUC with small variation and average extremely close to 1. Overall, MDA outperformed both LDA and QDA in both the two and three-class scenarios (Figures 8 & 9).
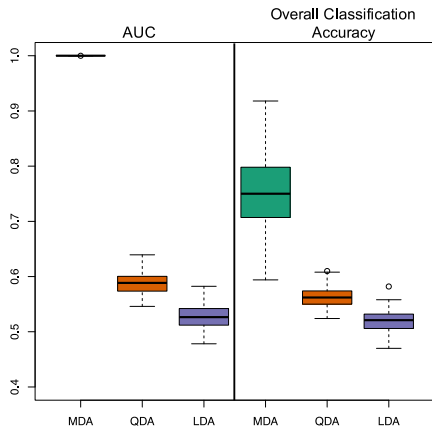
Figure 8: Simulation 4 two-class test set accuracies

Figure 9: Simulation 4 three-class test set accuracies
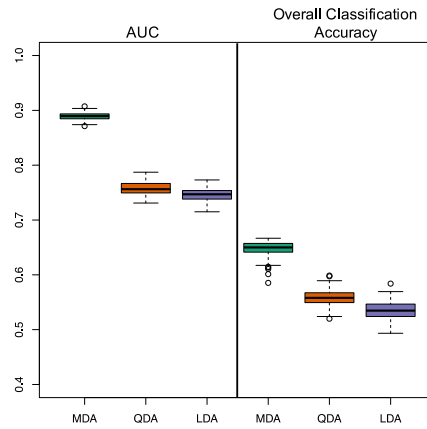
Prediction accuracies for all models increased as the covariation between the features increased (Figures 10 & 11). This was observed for both the two and three-class scenarios. MDA resulted in a steep increase in overall classification accuracy in the three-class setting. Furthermore, MDA had the highest average AUC at all levels of covariation and in both class settings.

18

Figure 10: Simulation 5 two-class
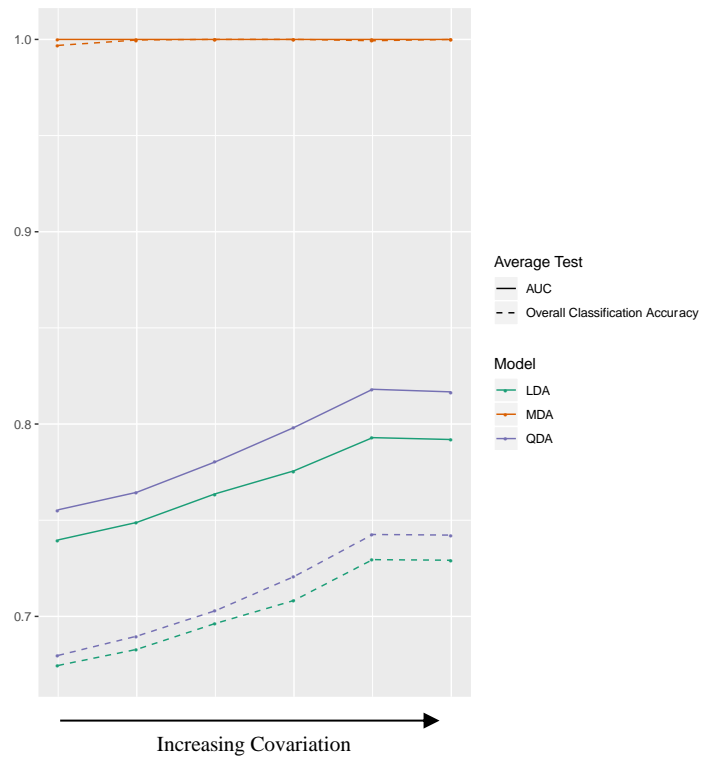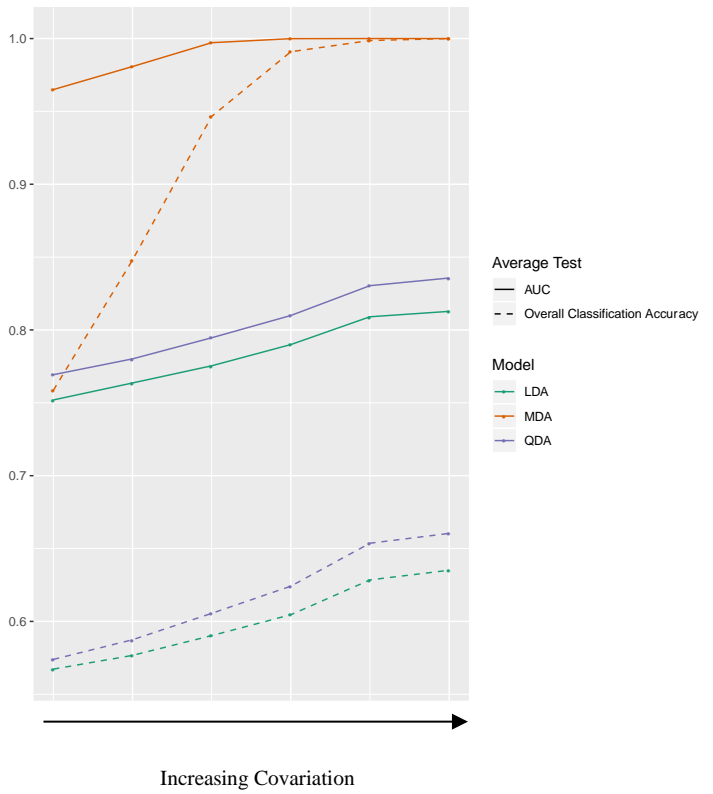average test set accuracies as
covariation increases

Increasing Covariation

Figure 11: Simulation 5 three-class
average test set accuracies as
covariation increases

Increasing Covariation

Cervical cancer quantitative cytology data results

The model was trained on 40% of the data and parameters were tuned on a validation set (30% of the data). The regularization parameter, α, in (2-3) and (2-4) was 0.2 in the two-class setting and 0.1 in both the three and four class setting. Table 3 provides results when patients were assigned to two histologic diagnosis classes. Overall classification accuracy, sensitivity, and specificity were all calculated using a threshold of 0.5. LDA had the highest overall classification accuracy and specificity, while MDA had the highest AUC and sensitivity.

Table 3: Cervical cancer two-class results on test set

Model	AUC	Overall Classification Accuracy	Sensitivity	Specificity	Q*-point (specificity, sensitivity)
MDA	0.7817	0.8215	0.3933	0.9097	(0.718, 0.764)
QDA	0.7125	0.8599	0.2809	0.9792	(0.775, 0.663)
LDA	0.7042	0.8637	0.2584	0.9884	(0.727, 0.652)

All three models performed similarly in the three and four class settings (Table 4 and 5). MDA had a slight advantage over LDA and QDA in the three-class scenario as the average pairwise AUC was approximately 0.05 higher.

Table 4: Cervical cancer three-class results on test set

Model	Average Pairwise AUC	Overall Classification Accuracy
MDA	0.6909	0.5374
QDA	0.6485	0.5432
LDA	0.6436	0.5374

Table 5: Cervical cancer four-class results on test set

Model	Average Pairwise AUC	Overall Classification Accuracy
MDA	0.6391	0.5374
QDA	0.6423	0.5374
LDA	0.6228	0.5374

DISCUSSION

Through comparisons using both cervical quantitative cytology data and simulated data sets, we have demonstrated that macrolevel discriminant analysis is a promising classification tool for nested data structures. Furthermore, MDA can be expanded to multi-class classification problems and is a better suited model than LDA or QDA under certain conditions.

Although several non-parametric classification methods exist that can handle non-normal data and data with heterogenous group structures, there has been limited research on methods that are robust to these deviations for nested data structures[13]. Our study of performance on simulated data sets indicates that MDA is robust to departures from typical assumptions required for LDA and QDA methods to perform well. Results from simulation 4 are particularly intriguing as MDA still performs exceptionally well even when the microlevel observations are sampled from two different gaussian distributions. Several real-life data sets exist that mimic this same pattern. For example, in quantitative cytology, a patient may have several cells containing features that look benign and some that resemble cancerous cells. MDA has the benefit of accounting for within patient heterogeneity and thus

21

more accurately classifies patients to diseased versus non-diseased groups. In LDA or QDA, researchers have typically calculated the summary statistics (e.g., mean) of each patients' cells and then proceeded to classification. MDA has an advantage over these traditional methods because a central measure of tendency may not be representative of a patients' true classification.

A large variation of overall classification accuracy on test sets using the MDA model was observed in a few simulations. This is likely because a constant regularization parameter was used for each simulation and not tuned to each specific training/test set. However, AUC calculations were stable and is a more accurate representation of performance primarily because it does not require adjustment of classification thresholds[13].

MDA had a slight advantage over LDA and QDA when predicting cervical histologic diagnosis classes as AUC was highest in the two and three class scenarios. However, we did not see a drastic improvement in predictive capabilities of MDA over LDA and QDA with application to the cervical quantitative data available. It is possible that a macrolevel patient effect did not exist in this data set, and further investigations should be performed to understand the underlying structure of the data. Additionally, it is promising to find that MDA performed as well as, or slightly better than, traditional discriminant analysis methods when separating into more defined classes even if the data is not well suited for this type of model.

A few limitations should be discussed in the implementation of MDA. One limitation is the long computation time required to build and take the inverse of the variance-covariance

matrix, $V_X^{j,k}$. This matrix has dimension $p \times n_j$ and is restructured for each patient for class prediction. Another possible limitation is the need for regularization of the within and between patient covariance matrices due to high dimensionality. Other regularization techniques, and procedures for handling high-dimensionality due to the large number of observations obtained for each patient, should be explored. This problem would be exacerbated as the number of measured features increases. If able to tackle high-dimensionality, prediction accuracies may improve as more important features are included when fitting the model.

Overall, MDA performs as well as, or better than, traditional discriminant analysis methods and makes use of all information provided at the micro and macro level. Our research included classification of up to four-classes for cervical cytology data and up to three-classes for simulated data. Most biomedical data do not require classification beyond four or five classes, but future research should explore prediction capabilities of MDA as the number of classes increases beyond four. Lastly, the MDA method should be applied to several other real-life repeated measures and nested data sets to determine its applicability to a wide array of environments.

REFERENCES

1. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International journal of cancer*. 2015;136(5):E359-E386.

2. Cancer facts & figures 2019. *American Cancer Society*. 2019.

3. Lix L, Sajobi T. Discriminant analysis for repeated measures data: A review. *Frontiers in psychology*. 2010;1:146.

4. Stoler MH, Schiffman M. Interobserver reproducibility of cervical cytologic and histologic interpretations: Realistic estimates from the ASCUS-LSIL triage study. *JAMA*. 2001;285(11):1500-1505.

5. Cox JT, Schiffman M, Solomon D. Prospective follow-up suggests similar risk of subsequent cervical intraepithelial neoplasia grade 2 or 3 among women with cervical intraepithelial neoplasia grade 1 or negative colposcopy and directed biopsy. *Obstet Gynecol*. 2003;188(6):1406-1412.

6. Kitchener HC, Castle PE, Cox JT. Achievements and limitations of cervical cytology screening. *Vaccine*. 2006;24:S63-S70.

7. Pinto NR, Applebaum MA, Volchenboum SL, et al. Advances in risk classification and treatment strategies for neuroblastoma. *Journal of clinical oncology*. 2015;33(27):3008.

8. Bingenheimer JB, Raudenbush SW. Statistical and substantive inferences in public health: Issues in the application of multilevel models. *Annu.Rev.Public Health*. 2004;25:53-77.

9. Guo Y, Hastie T, Tibshirani R. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*. 2007;8(1):86-100.

10. Yamal J, Follen M, Guillaud M, Cox DD. Classifying tissue samples from measurements on cells with within-class tissue sample heterogeneity. *Biostatistics*. 2011;12(4):695-709.

11. Guillaud M, Benedet JL, Cantor SB, Staerkel G, Follen M, MacAulay C. DNA ploidy compared with human papilloma virus testing (hybrid capture II) and conventional cervical cytology as a primary screening test for cervical high-grade lesions and cancer in 1555 patients with biopsy confirmation. *Cancer*. 2006;107(2):309-318.

12. James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning.* New York: Springer; 2013.

13. Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Mach Learning*. 2001;45(2):171-186.