PERFORMANCE OF LINEAR DISCRIMINANT ANALYSIS IN STOCHASTIC SETTINGS

Amin Zollanvari[1], *Jianping Hua*[2], *Edward R. Dougherty*[1,2]

[1]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX
[2]Translational Genomics Research Institute (TGEN), Phoenix, AZ

ABSTRACT

This paper provides, for the first time, exact analytical expressions for the first moment of the true error of linear discriminant analysis (LDA) when the data are univariate and taken from two stochastic Gaussian processes. We assume a general setting in which the sample data from each class do not need to be identically distributed or independent within or between classes. As an application of this framework, we characterize the performance of LDA in situations that the data are generated from autoregressive models of the first order.

Index Terms— Linear discriminant analysis, Stochastic settings, Non-i.i.d data, Expected error, Gaussian processes

1. INTRODUCTION

It is common in practice to assume that the training data used to construct a classifier are independent and identically distributed (i.i.d). Should the data be dependent or not identically distributed, the classifier performance is affected. This paper presents a mathematical framework for analytically studying classifiers in such situations in general, and the univariate LDA (linear discriminant analysis) classifier in particular. We pay particular attention to the univariate LDA model because it is possible to obtain closed-form (not asymptotic) results for moments of the error – in analogy to moments for the error [1, 2] and error estimates [1, 3] for univariate LDA with i.i.d. sampling. The desired framework is achieved by placing classifier performance in a stochastic setting where the training data are univariate dependent and not necessarily identically distributed (non-i.i.d.).

Motivation for this line of research goes back to the early 1970's when Basu and Odell observed in remote sensing applications that the conditional expected true error of LDA is commonly higher than what is expected from a theoretical analysis [4]. They associated this observation with violation of the independence assumption on the training data. Following this work, several researchers obtained asymptotic expressions for the first moment of LDA true error in situations that the data have a specific correlation structure [5, 6].

Typically, large-sample asymptotic results are not helpful in small-sample situations [7]. This understanding led us to study the distribution and exact moments of LDA true error and comnon estimators [3, 7, 8]. Having laid the groundwork for analyzing LDA related statistics in small-sample situations, in this work, we establish a stochastic framework for studying univariate LDA true error in a general setting. We neither impose a specific correlation structure on the training data, nor do we assume the training data have necessarily the same mean or variance. For example the basic assumption in [4, 5, 6] is that the training data of the two classes are taken separately from two class conditional densities Π_0, for class 0, and Π_1, for class 1. This assumption immediately imposes several restrictions on the problem: the training data from each class have the same mean and variance (because they are coming from the same distribution) and, furthermore, only intraclass correlations exist. The stochastic setting permits us to generalize such assumptions to training data being correlated across classes or the samples from each class being differently distributed. To model such data we employ Gaussian processes and we assume the samples are taken from class conditional processes rather than class conditional densities.

2. LINEAR DISCRIMINANT ANALYSIS AND ERROR ESTIMATION: INDEPENDENT SAMPLING

In this section, we present the traditional sampling scenario in which LDA is employed. Consider a set of $n = n_0 + n_1$ independent samples in R^p, where $X_1, X_2, \ldots, X_{n_0}$ come from population Π_0 and $X_{n_0+1}, X_{n_0+2}, \ldots, X_{n_0+n_1}$ coming from population Π_1, with p being the dimensionality of each sample. Population Π_i is assumed to follow a univariate Gaussian distribution $N(\mu_i, \sigma_i^2)$, for $i = 0, 1$. In general, *Linear Discriminant Analysis* (LDA) utilizes the Anderson W statistic

$$W(\bar{X}^0, \bar{X}^1, X) = \left(X - \frac{\bar{X}^0 + \bar{X}^1}{2} \right)^T \hat{\mathbf{\Sigma}}^{-1} \left(\bar{X}^0 - \bar{X}^1 \right), \tag{1}$$

where $\bar{X}^0 = \frac{1}{n_0} \sum_{i=1}^{n_0} X_i$ and $\bar{X}^1 = \frac{1}{n_1} \sum_{i=n_0+1}^{n_0+n_1} X_i$ are the sample means for each class and $\hat{\mathbf{\Sigma}}$ is the pooled estimate of the covariance matrix, which is assumed to be common in the LDA discriminant. Given $\bar{X}^0$ and $\bar{X}^1$, the designed LDA classifier is given by

$$\psi(X) = \begin{cases} 1, & \text{if } W(\bar{X}^0, \bar{X}^1, X) \leq c \\ 0, & \text{if } W(\bar{X}^0, \bar{X}^1, X) > c \end{cases}, \tag{2}$$

with c being a constant. It is commonly assumed that c is zero, which is the assumption we also make throughout this paper. Therefore, the sign of W determines the classification of the sample point X. In the univariate model, $\hat{\mathbf{\Sigma}}$ reduces to $\hat{\sigma}^2$, and (1) reduces to $W(\bar{X}^0, \bar{X}^1, X) = \frac{1}{\hat{\sigma}^2}(X - \bar{X})\left(\bar{X}^0 - \bar{X}^1\right)$. Since $\hat{\sigma}^2 \geq 0$ this reduce to

$$W(\bar{X}^0, \bar{X}^1, X) = (X - \bar{X})\left(\bar{X}^0 - \bar{X}^1\right) \tag{3}$$

where $\bar{X} = \frac{\bar{X}^0 + \bar{X}^1}{2}$. Given the training data S_n (and thus $\bar{X}_0$ and $\bar{X}_1$), the classification error is given by

$$\begin{aligned}\epsilon = & P(W(\bar{X}^0, \bar{X}^1, X) \leq 0, X \in \Pi_0 | \bar{X}^0, \bar{X}^1) \\ & + P(W(\bar{X}^0, \bar{X}^1, X) > 0, X \in \Pi_1 | \bar{X}^0, \bar{X}^1) = \alpha_0 \epsilon^0 + \alpha_1 \epsilon^1\end{aligned} \tag{4}$$

where $\alpha_i = P(X \in \Pi_i)$ is the a priori mixing probability for population Π_i and ϵ^i is the error rate specific to population Π_i, with

$$\epsilon^i = P((-1)^i W(\bar{X}^0, \bar{X}^1, X) \leq 0 | X \in \Pi_i, \bar{X}^0, \bar{X}^1). \tag{5}$$

The first moment of the actual error is given by

$$E[\epsilon] = \sum_{i=0}^{1} \alpha_i P((-1)^i W(\bar{X}^0, \bar{X}^1, X) \leq 0 | X \in \Pi_i) \tag{6}$$

3. PERFORMANCE OF LDA CLASSIFIER IN UNIVARIATE GAUSSIAN DEPENDENT SAMPLING (UGDS) MODEL OF BINARY CLASSIFICATION

We now provide the mathematical framework to study LDA true error in a stochastic setting, thereby allowing us to study, for the first time, the effect of having non-i.i.d. data on LDA performance.

Definition 1 : A process $\mathbf{X}_t = \{X_t : t \in \mathbf{T}\}$ with $\mathbf{T}$ being an ordered set, is called a Gaussian process if any finite-dimensional vector $[X_{t_1}, X_{t_2}, ..., X_{t_n}]^T$ has the multivariate normal distribution $N(\boldsymbol{\mu}_t, \mathbf{\Sigma}_t)$, where $\boldsymbol{\mu}_t = [E(X_{t_1}), E(X_{t_2}), ..., E(X_{t_n})] = [\mu_1, \mu_2, ..., \mu_n]$ and $\mathbf{\Sigma}_t$ is the covariance matrix dependent on $T = [t_1, t_2, ..., t_n]$. For the ease of notations we omit the subscript t from $\boldsymbol{\mu}_t$ and $\mathbf{\Sigma}_t$.

Definition 2 : We refer to the following sampling procedure as the *Univariate Gaussian Dependent Sampling (UGDS) Model of Binary Classification*: $\mathbf{X}_t^i = \{X_{t^i}^i : t^i \in \mathbf{T}^i\}$, with $\mathbf{T}^i$ being two ordered sets for $i = 0, 1$, are two Gaussian processes such that any finite-dimensional vector constructed by stacking the random variables of $\mathbf{X}_{t^i}^i$ as $[X_{t_1^0}^0, X_{t_2^0}^0, ..., X_{t_{n_0}^0}^0, X_{t_1^1}^1, X_{t_2^1}^1, ..., X_{t_{n_1}^1}^1]^T$ possesses a multivariate normal distribution $N(\boldsymbol{\mu}, \mathbf{\Sigma})$, where $\boldsymbol{\mu} = [\mu_1^0, \mu_2^0, ..., \mu_{n_0}^0, \mu_1^1, \mu_2^1, ..., \mu_{n_1}^1]$, and

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{n_0 \times n_0}^{00} & \mathbf{\Sigma}_{n_0 \times n_1}^{01} \\ \mathbf{\Sigma}_{n_1 \times n_0}^{10} & \mathbf{\Sigma}_{n_1 \times n_1}^{11} \end{bmatrix} \tag{7}$$

is a positive definite covariance matrix.

This model is univariate because both processes, $\mathbf{X}_{t^0}^0$ and $\mathbf{X}_{t^1}^1$, are collections of univariate random variables, not necessarily with the same means or variances. $\mathbf{X}_{t^0}^0$ and $\mathbf{X}_{t^1}^1$ are called class conditional processes. For ease of notations and without loss of mathematical generality, we assume that $\mathbf{T}^0$ and $\mathbf{T}^1$ are the same set and, therefore, we omit the superscript i from t^i. Thus, henceforth we denote $\mathbf{X}_{t^i}^i$ by $\mathbf{X}_t^i$ and the stacked vector $[X_{t_1^0}^0, X_{t_2^0}^0, ..., X_{t_{n_0}^0}^0, X_{t_1^1}^1, X_{t_2^1}^1, ..., X_{t_{n_1}^1}^1]^T$ by $[X_{t_1}^0, X_{t_2}^0, ..., X_{t_{n_0}}^0, X_{t_1}^1, X_{t_2}^1, ..., X_{t_{n_1}}^1]^T$.

Employing LDA with the UGDS model instead of traditional independent sampling, the LDA rule becomes

$$W(\bar{X}_T^0, \bar{X}_T^1, X_t) = \left(X_t - \frac{\bar{X}_T^0 + \bar{X}_T^1}{2}\right)^T \hat{\mathbf{\Sigma}}_T^{-1}\left(\bar{X}_T^0 - \bar{X}_T^1\right), \tag{8}$$

where $\bar{X}_T^0 = \frac{1}{n_0}\sum_{i=1}^{n_0} X_{t_i}^0$ and $\bar{X}_t^1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{t_i}^1$ are the sample means for each class and $\hat{\mathbf{\Sigma}}_T$ is the common pooled estimate of the covariance matrix of the classes. Similar to (3), in the univariate case W reduces to

$$W(\bar{X}_T^0, \bar{X}_T^1, X_t) = (X_t - \bar{X}_T)\left(\bar{X}_T^0 - \bar{X}_T^1\right), \tag{9}$$

where $\bar{X}_T = \frac{\bar{X}_T^0 + \bar{X}_T^1}{2}$. The designed LDA classifier is given by

$$\psi(X_{t_s}) = \begin{cases} 1, & \text{if } W(\bar{X}_T^0, \bar{X}_T^1, X_{t_s}) \leq 0 \\ 0, & \text{if } W(\bar{X}_T^0, \bar{X}_T^1, X_{t_s}) > 0 \end{cases}. \tag{10}$$

3.1. Stochastic true error and its moments

We denote a future test sample by $X_{t_s}^i$, where i indicates the class conditional process in which the sample is coming from, i.e. either $\mathbf{X}_t^0$ or $\mathbf{X}_t^1$. The auto-covariance sequence of $X_{t_s}^i$ with the training data is defined as

$$\rho_s^{ik}(j) = E[(X_{t_s}^i - \mu_s^i)(X_{t_j}^k - \mu_j^k)], i, k = 0, 1,\ j = 1, 2, ..., n_k \tag{11}$$

where $\rho_s^{ik}(j)$ is the j^{th} element of the sequence ρ_s^{ik}. Since $X_{t_s}^i$ is a future sample, we assume $2 \leq \max\{n_0, n_1\} < s$, unless otherwise stated. Throughout the paper, we use S_A to denote the sum of all elements of a matrix or vector A. For instance, $S_{\rho_s^{ik}} = \sum_{j=1}^{n_i} \rho_s^{ik}(j)$.

The true classifier error under the UGDS model is a function of t_s. Sample points at t_s can come from either processes and the classifier may misclassify any of these. Hence,

$$\epsilon_{t_s} = \alpha_{t_s}^0\ \epsilon_{t_s}^0 + \alpha_{t_s}^1\ \epsilon_{t_s}^1. \tag{12}$$

where $\alpha_{t_s}^i = P(X_{t_s} \in \mathbf{X}_t^i), i = 0, 1$, is the a priori mixing probability of the two processes $\mathbf{X}_t^0$ and $\mathbf{X}_t^1$ at t_s and $\epsilon_{t_s}^i$ is the error rate specific to each process, with

$$\epsilon_{t_s}^i = P((-1)^i W(\bar{X}_T^0, \bar{X}_T^1, X_{t_s}) \leq 0 | \bar{X}_T^0, \bar{X}_T^1, X_{t_s} \in \mathbf{X}_t^i) \tag{13}$$

By replacing $W(\bar{X}_T^0, \bar{X}_T^1, X_{t_s})$ with any proper statistic used in other classifiers, this stochastic definition of true error applies to other rules. The expected performance of true error is also specific to t_s and is then defined to be:

$$E[\epsilon_{t_s}] = \sum_{i=0}^{1} \alpha_{t_s}^i P((-1)^i W(\bar{X}_T^0, \bar{X}_T^1, X_{t_s}) \leq 0 | X_{t_s} \in \mathbf{X}_t^i)$$

3.2. Expected performance of LDA in UGDS model

The first moment of the classification error for LDA under the UGDS model is expressed exactly according to the following theorem, where $Z = (z_1, z_2)^T < 0$ means $z_1 < 0, z_2 < 0$.

Theorem 1 *Under UGDS model, the expected true error of LDA at t_s is*

$$\begin{aligned} E[\epsilon^D_{t_s, n_0+n_1}] &= \alpha_{t_s}^0 \left[P(Z_s^I < 0) + P(Z_s^I \geq 0) \right] \\ &+ \alpha_{t_s}^1 \left[P(Z_s^{II} < 0) + P(Z_s^{II} \geq 0) \right], \end{aligned} \tag{14}$$

where $Z_{t_s}^I$ and $Z_{t_s}^{II}$ are Gaussian bivariate vectors with:

$$\mu_{Z_s^I} = \begin{bmatrix} \mu_s^0 - \frac{\bar{\mu}}{2} & -\mu' \end{bmatrix}^T, \quad \mu_{Z_s^{II}} = \begin{bmatrix} \mu_s^1 - \frac{\bar{\mu}}{2} & \mu' \end{bmatrix}^T$$

$$\mathbf{\Sigma}_{Z_s^I} = \begin{bmatrix} (\sigma_s^0)^2 - \frac{S_{\rho_s^{00}}}{n_0} - \frac{S_{\rho_s^{01}}}{n_1} + \frac{S_{\mathbf{\Sigma}^{00}}}{4n_0^2} + \frac{S_{\mathbf{\Sigma}^{11}}}{4n_1^2} + \frac{S_{\mathbf{\Sigma}^{01}}}{2n_0 n_1} & \frac{-S_{\rho_s^{00}}}{n_0} + \frac{S_{\rho_s^{01}}}{n_1} + \frac{S_{\mathbf{\Sigma}^{00}}}{2n_0^2} - \frac{S_{\mathbf{\Sigma}^{11}}}{2n_1^2} \\ . & \frac{S_{\mathbf{\Sigma}^{00}}}{n_0^2} + \frac{S_{\mathbf{\Sigma}^{11}}}{n_1^2} - \frac{2S_{\mathbf{\Sigma}^{01}}}{n_0 n_1} \end{bmatrix}$$

$$\mathbf{\Sigma}_{Z_s^{II}} = \begin{bmatrix} (\sigma_s^1)^2 - \frac{S_{\rho_s^{11}}}{n_1} - \frac{S_{\rho_s^{10}}}{n_0} + \frac{S_{\mathbf{\Sigma}^{00}}}{4n_0^2} + \frac{S_{\mathbf{\Sigma}^{11}}}{4n_1^2} + \frac{S_{\mathbf{\Sigma}^{01}}}{2n_0 n_1} & \frac{-S_{\rho_s^{11}}}{n_1} + \frac{S_{\rho_s^{10}}}{n_0} - \frac{S_{\mathbf{\Sigma}^{00}}}{2n_0^2} + \frac{S_{\mathbf{\Sigma}^{11}}}{2n_1^2} \\ . & \frac{S_{\mathbf{\Sigma}^{00}}}{n_0^2} + \frac{S_{\mathbf{\Sigma}^{11}}}{n_1^2} - \frac{2S_{\mathbf{\Sigma}^{01}}}{n_0 n_1} \end{bmatrix} \tag{15}$$

where $\bar{\mu} = \frac{\sum_{i=1}^{n_0} \mu_i^0}{n_0} + \frac{\sum_{i=1}^{n_1} \mu_i^1}{n_1}$, $\mu' = \frac{\sum_{i=1}^{n_0} \mu_i^0}{n_0} - \frac{\sum_{i=1}^{n_1} \mu_i^1}{n_1}$, and μ_s^i and $(\sigma_s^i)^2$ are the mean and variance of random variables at t_s from class $i, i = 0, 1$, with the auto-covariance ρ_s^{ik} defined as in (11).

Proof. From (9), it follows that

$$E[\epsilon_{t_s}^0] = P(W(\bar{X}_T^0, \bar{X}_T^1, X_{t_s}) \leq 0 | X_{t_s} \in \mathbf{X}_t^0) =$$
$$P(X_{t_s} - \bar{X}_T < 0, \bar{X}_T^0 - \bar{X}_T^1 > 0) + P(X_{t_s} - \bar{X}_T \geq 0, \bar{X}_T^0 - \bar{X}_T^1 < 0)$$

where $\bar{X}_T = \frac{\bar{X}_T^0 + \bar{X}_T^1}{2}$. Expanding $\bar{X}_T^0$ and $\bar{X}_T^1$ as $\frac{1}{n_0} \sum_{i=1}^{n_0} X_{t_i}^0$ and $\frac{1}{n_1} \sum_{i=1}^{n_1} X_{t_i}^1$ results in

$$E[\epsilon_{t_s}^0] = P(Z_s^{\mathrm{I}} < 0) + P(Z_s^{\mathrm{I}} \geq 0) \tag{16}$$

where $Z_s^{\mathrm{I}} = AY_s^0$, and $Y_s^0 = [X_{t_s}^0, \ldots, X_{t_{n_0}}^0, X_{t_1}^1, \ldots, X_{t_{n_1}}^1]^T$, where the super index 0 in $X_{t_s}^0$ is to denote explicitly $X_{t_s} \in \mathbf{X}_t^0$, and

$$A = \begin{bmatrix} 1 & \frac{-1}{2n_0} & \frac{-1}{2n_0} & \cdots & \frac{-1}{2n_0} & \frac{-1}{2n_1} & \cdots & \frac{-1}{2n_1} \\ 0 & \frac{-1}{n_0} & \frac{-1}{n_0} & \cdots & \frac{-1}{n_0} & \frac{1}{n_1} & \cdots & \frac{1}{n_1} \end{bmatrix} \tag{17}$$

Therefore, Z_s^{I} is a Gaussian random vector with mean $A\mu_{Y_s^0}$ and covariance matrix $A\mathbf{\Sigma}_{Y_s^0}A^T$. Plugging in the values of $\mu_{Y_s^0} = [\mu_s^0, \mu_1^0, \mu_2^0, ..., \mu_{n_0}^0, \mu_1^1, ..., \mu_{n_1}^1]$ and noting the fact that the j^{th} element of vector ρ_s^{ik} is defined as $\rho_s^{ik}(j) = E[(X_{t_s}^i - \mu_s^i)(X_{t_j}^k - \mu_j^k)]$, $i, k = 0, 1$, $j = 1, 2, ..., n_k$, leads to the expression stated in Theorem 1. Evaluating the mean and covariance matrix of vector Z_s^{II}, which is the counterpart for $E[\epsilon_{t_s}^1]$ is entirely similar, by considering $P(W(\bar{X}_T^0, \bar{X}_T^1, X_{t_s}) > 0 | \bar{X}_T^0, \bar{X}_T^1, X_{t_s} \in \mathbf{X}_t^1)$.■

As an application of Theorem 1, we assume that the training data of the two classes are taken separately from two independent autoregressive models of first order, AR(1), as follows:

$$X_t^i = c_i + \psi_i X_{t-1}^i + Z_t^i, \quad i = 0, 1 \tag{18}$$

where ψ_i is a constants such that $0 < |\psi_i| < 1, i = 0, 1$, and $Z_t^0 \sim N(0, \sigma_0^2)$ and $Z_t^1 \sim N(0, \sigma_1^2)$ for all t and are independent from each other. Then $\mathbf{X}_t^0 = \{X_t^0 : 0 < t < \infty\}$ and $\mathbf{X}_t^1 = \{X_t^1 : 0 < t < \infty\}$ are two independent covariance-stationary processes and we have the following theorem.

Theorem 2 *Let $\mathbf{X}_t^0$, $\mathbf{X}_t^1$ in the UGDS model be defined by the two independent covariance-stationary AR(1) processes as defined in (18). Then the expected true error of LDA constructed using the training samples $[X_{t_1}^0, X_{t_2}^0, ..., X_{t_{n_0}}^0]$ and $[X_{t_1}^1, X_{t_2}^1, ..., X_{t_{n_1}}^1]$ at t_s, where max$\{n_0, n_1\} < s$, is*

$$\begin{aligned} E[\epsilon^{AR(1)}_{t_s, n_0+n_1}] &= \alpha_{t_s}^0 \left[P(Z_s^I < 0) + P(Z_s^I \geq 0) \right] \\ &+ \alpha_{t_s}^1 \left[P(Z_s^{II} < 0) + P(Z_s^{II} \geq 0) \right], \end{aligned} \tag{19}$$

where $Z_{t_s}^I$ and $Z_{t_s}^{II}$ are Gaussian bivariate vectors with

$$\mu_{Z_s^I} = \begin{bmatrix} \frac{\mu}{2} & -\mu \end{bmatrix}^T, \quad \mu_{Z_s^{II}} = \begin{bmatrix} \frac{-\mu}{2} & \mu \end{bmatrix}^T$$

$$\mathbf{\Sigma}_{Z_s^I} = \begin{bmatrix} \frac{\sigma_0^2}{1-\psi_0^2} - \frac{S_{\rho_s^{00}}}{n_0} + \frac{S_{\mathbf{\Sigma}^{00}}}{4n_0^2} + \frac{S_{\mathbf{\Sigma}^{11}}}{4n_1^2} & \frac{-S_{\rho_s^{00}}}{n_0} + \frac{S_{\mathbf{\Sigma}^{00}}}{2n_0^2} - \frac{S_{\mathbf{\Sigma}^{11}}}{2n_1^2} \\ . & \frac{S_{\mathbf{\Sigma}^{00}}}{n_0^2} + \frac{S_{\mathbf{\Sigma}^{11}}}{n_1^2} \end{bmatrix}$$

$$\mathbf{\Sigma}_{Z_s^{II}} = \begin{bmatrix} \frac{\sigma_1^2}{1-\psi_1^2} - \frac{S_{\rho_s^{11}}}{n_1} + \frac{S_{\mathbf{\Sigma}^{00}}}{4n_0^2} + \frac{S_{\mathbf{\Sigma}^{11}}}{4n_1^2} & \frac{-S_{\rho_s^{11}}}{n_1} - \frac{S_{\mathbf{\Sigma}^{00}}}{2n_0^2} + \frac{S_{\mathbf{\Sigma}^{11}}}{2n_1^2} \\ . & \frac{S_{\mathbf{\Sigma}^{00}}}{n_0^2} + \frac{S_{\mathbf{\Sigma}^{11}}}{n_1^2} \end{bmatrix} \tag{20}$$

where

$$\mu = \frac{c_0}{1-\psi_0} - \frac{c_1}{1-\psi_1}, S_{\rho_s^{ii}} = \frac{\psi_i^{(s-n_i)} \sigma_i^2}{1-\psi_i^2} \left(\frac{1-\psi_i^{n_i}}{1-\psi_i} \right),$$
$$S_{\mathbf{\Sigma}^{ii}} = \frac{\sigma_i^2}{(1-\psi_i^2)(1-\psi_i)} \left[n_i(1+\psi_i) - 2\psi_i \left(\frac{1-\psi_i^{n_i}}{1-\psi_i} \right) \right]. \tag{21}$$

Proof. Since the Z_t^i's are Gaussian, X_t^0 and X_t^1 are covariance-stationary [9] and the vectors $X_{n_0}^0 = [X_{t_1}^0, X_{t_2}^0, ..., X_{t_{n_0}}^0]^T$ and $X_{n_0}^0 = [X_{t_1}^1, X_{t_2}^1, ..., X_{t_{n_1}}^1]^T$ are distributed normally as

$X^i_{n_i} \sim N(\boldsymbol{\mu}^i, \boldsymbol{\Sigma}^i),\ i = 0, 1$:

$$\begin{aligned} &\boldsymbol{\mu}^i = [\mu^i, \mu^i, ..., \mu^i]^T_{1\times n_i}, \boldsymbol{\Sigma}^i(k,l) = \frac{\psi_i^{|k-l|}}{1-\psi_i^2}\sigma_i^2, \\ &\rho_s^{ii}(k) = \frac{\psi_i^{s-k}}{1-\psi_i^2}\sigma_i^2,\ \rho_s^{01} = \mathbf{0}_{1\times n_1},\ \rho_s^{10} = \mathbf{0}_{1\times n_0}, \end{aligned} \tag{22}$$

where $k, l = 1, 2, ..., n_i$, $\mu_i = \frac{c_i}{1-\psi_i}$, and $\boldsymbol{\Sigma}^i(k,l)$ denotes the entry in the k^{th} row and l^{th} column of matrix $\boldsymbol{\Sigma}^i$. The result follows by replacing (22) in Theorem 1. ■

4. NUMERICAL EXAMPLE

Experiment 1: In this experiment, we consider the data are generated by the first order autoregressive model defined in (18). We assume $\alpha^0_{t_s} = \alpha^1_{t_s}$, $n_0 = n_1 = n$, $\sigma_0 = \sigma_1 = 1$, and $\psi_0 = \psi_1 = \psi \in [-0.95, 0.95]$. We consider various cases where $c_0 = 0.5,\ 0.75,\ 1,\ 1.5$ with $c_0 = -c_1$. Figure 1 shows the exact expectation of LDA true error constructed on the data coming from such experiment. These results are exact and are calculated from Theorem 2. The figure suggests that increasing ψ decreases $E[\epsilon^{AR(1)_\psi}_{t_s,2n}]$ and therefore, $E[\epsilon^{AR(1)_\psi}_{t_s,2n}]$ seems to be a decreasing function of ψ in this experiment. Furthermore, the figure suggests that $E[\epsilon^{AR(1)_\psi}_{t_s,2n}] < E[\epsilon^{I}_{t_s,2n}]$ for $0 < \psi < 1$ and $E[\epsilon^{AR(1)_\psi}_{t_s,2n}] > E[\epsilon^{I}_{t_s,2n}]$ for $0 < \psi < 1$, in which $E[\epsilon^{I}_{t_s,2n}] = E[\epsilon^{AR(1)_{\psi=0}}_{t_s,2n}]$, i.e. $E[\epsilon^{I}_{t_s,2n}]$ is LDA expected error where we have independent data. Currently, we are analytically investigating the behavior of $E[\epsilon^{AR(1)_\psi}_{t_s,2n}]$ as a function of ψ.

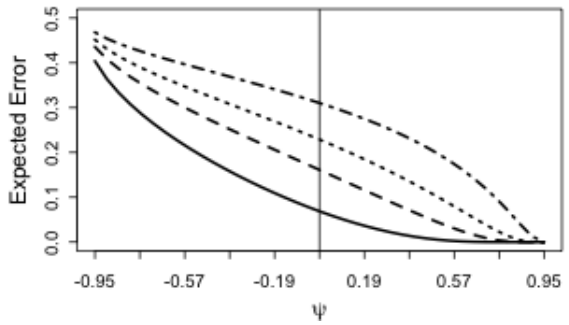

Fig. 1. Exact expectation of LDA true error of the first-order autoregressive model in the Experiment as a function of $\psi := \psi_0 = \psi_1$ for $n_0 = n_1 = 25$ and $s - n_0 = 10$; Plot keys: solid := $c_0 = 1.5$; dash := $c_0 = 1$; dot := $c_0 = 0.75$; dash-dot := $c_0 = 0.5$. The cross section of each curve with the vertical solid line in each plot shows the magnitude of the expectation for i.i.d sampling situation for the corresponding scenario.

Experiment 2: This experiment is an example derived from gene-expression data used in studying the prognosis of breast-cancer using 70 genes with high prognostic ability [10]. Following [11], we divide the 307 individuals used in this study into 64 "poor" prognosis (class 0) versus 243 "good" prognosis (class 1) patients. The gene expression data used in this study have been collected by triplicating each gene on each microarray and then duplicating each measurement by dye-swaping. Therefore, for each patient, each gene, we have six measurements, three of which are positively correlated with themselves and negatively correlated with others. We consider a scenario in which the experimenter is only given six measurements taken from one patient from class 0 and six measurements from another patient from class 1, and a univariate LDA classifier is desired to differentiate the two groups using ALDH4 gene, which has the highest correlation with prognosis of breast cancer [11]. Therefore, in this scenario, the experimenter is given 12 "technical" replicates in total, which are now treated as our "sample points". To verify the Gaussianity of each of the 12 random variables that are used in this example, i.e. $[X^0_{t_1}, X^0_{t_2}, ..., X^0_{t_6}, X^1_{t_1}, X^1_{t_2}, ..., X^1_{t_6}]^T$, a Shapiro-Wilk test is applied on the full dataset corresponding to each random variable. This test did not reject Gaussianity of the random variables over either of the classes at a 95% significance level after employing the Bonferroni correction of multi-hypothesis tests. Sample means, variances, and correlation, computed on the full dataset, were used as estimates of the unknown true means, variances, and the correlation structure between samples needed in Theorem 1. Using Theorem 1, the expected performance of a classifier, $E[\epsilon^D_{t_s,12}]$, to differentiate samples distributed as $X^0_{t_5}$ from samples distributed as $X^1_{t_1}$ is 0.475. To practically investigate this expected performance, we construct a classifier on each possible combination among $243 \times 64 = 15552$ combinations of 6 samples from either classes and each time we test the accuracy of the designed classifier on the $64 - 1 = 63$ remaining realizations of $X^0_{t_5}$ and $243 - 1 = 242$ realizations of $X^1_{t_1}$. The accuracy computed in this way is 0.479, which is almost the same as what is computed from Theorem 1. On the other hand, if in this experiment one ignores the correlation structure between samples, then from Theorem 1 in [3], the expected performance of LDA seems to be 0.374, which is not correct.

5. CONCLUSION

In many applications, the assumption of having i.i.d. training samples is violated. This paper characterizes, for the first time, the exact performance of univariate LDA classification in situations that the data are not independent or identically distributed. We achieved this by considering LDA in a stochastic setting in which the samples are taken from two class conditional Gaussian processes, which are not necessarily independent. The results show that the correlation structure of data can be either beneficial or detrimental in terms of classification performance.

6. REFERENCES

[1] M. Hills, "Allocation rules and their error rates," *J. Royal Statist. Soc. Ser. B (Methodological)*, vol. 28, pp. 1–31, 1966.

[2] M. J. Sorum, "Estimating the expected probability of misclassification for a rule based on the linear discriminant function: Univariate normal case," *Technometrics*, vol. 15, pp. 329–339, 1973.

[3] A. Zollanvari, U. M. Braga-Neto, and E. R. Dougherty, "Exact representation of the second-order moments for resubstitution and leave-one-out error estimation for linear discriminant analysis in the univariate heteroskedastic gaussian model," *Pattern Recogn.*, vol. 45, pp. 908–917, 2012.

[4] J. P. Basu and P. L. Odell, "Effect of intraclass correlation among training samples on the misclassification probabilities of bayes' procedure," *Pattern Recogn.*, vol. 6, pp. 13–16, 1974.

[5] G. J. McLachlan, "Further results on the effect of interclass correlation among training samples in discriminant analysis," *Pattern Recogn.*, vol. 8, pp. 273–275, 1976.

[6] C. R. O. Lawoko and G. J. McLachlan, "Discrimination with autocorrelated observations," *Pattern Recogn.*, vol. 18, pp. 145–149, 1985.

[7] A. Zollanvari, U. M. Braga-Neto, and E. R. Dougherty, "Joint sampling distribution between actual and estimated classification errors for linear discriminant analysis," *IEEE Trans. Inf. Theory*, vol. 56, pp. 784–804, 2010.

[8] ——, "Analytic study of performance of error estimators for linear discriminant analysis," *IEEE Trans. Sig. Proc.*, vol. 59, pp. 4238–4255, 2011.

[9] J. D. Hamilton, *Time Series Analysis*. NJ: Princeton University Press, 1994.

[10] M. Buyse, S. Loi, and et al., "Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer," *Journal of the National Cancer Institute*, vol. 98, pp. 1183–1192, 2006.

[11] L. van 't Veer, H. Dai, and et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530–6, 2002.