Published in final edited form as:
Pattern Recognit. 2013 November ; 46(11): 3017–3029. doi:10.1016/j.patcog.2013.04.002.

Analytical Study of Performance of Linear Discriminant Analysis in Stochastic Settings

Amin Zollanvari[a,b], **Jianping Hua**[c], and **Edward R. Dougherty**[a,c]

Amin Zollanvari: amin_zoll@neo.tamu.edu; Jianping Hua: jhua@tgen.org; Edward R. Dougherty: edward@ece.tamu.edu

[a]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843

[b]Department of Statistics, Texas A&M University, College Station, TX 77843

[c]Translational Genomics Research Institute (TGEN), Phoenix, AZ 85004

Abstract

This paper provides exact analytical expressions for the first and second moments of the true error for linear discriminant analysis (LDA) when the data are univariate and taken from two stochastic Gaussian processes. The key point is that 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. We compare the true errors of designed classifiers under the typical i.i.d. model and when the data are correlated, providing exact expressions and demonstrating that, depending on the covariance structure, correlated data can result in classifiers with either greater error or less error than when training with uncorrelated data. The general theory is applied to autoregressive and moving-average models of the first order, and it is demonstrated using real genomic data.

Keywords

Linear discriminant analysis; Stochastic settings; Correlated data; Non-i.i.d data; Expected error; Gaussian processes; Auto-regressive models; Moving-average models

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.

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.

Correspondence to: Amin Zollanvari, amin_zoll@neo.tamu.edu.

To study the effect of correlated training data on the performance of LDA, Basu and Odell [4] used numerical examples under an equicorrellated structure of samples (see Appendix for definition of various correlation structures). They showed that misclassification probabilities change under such structures. Afterwards, McLachlan [5] used asymptotic analysis to show that even under a simple-equicorrelated structure the probability of misclassification changes. Later, Tubbs [6] used a similar asymptotic analysis but with a serially correlated structure among training data. He considered further simplifying assumptions to show that the asymptotic error rate changes with serially correlated data having positive correlations. Lawoko and McLachlan [7] used the same serially correlated structure and obtained a different asymptotic expansion of LDA true error from the one that Tubbs previously achieved in [6]. This type of asymptotic analysis was later used in [7, 8] to characterize the asymptotic expected true error of univariate LDA and Z-statistics assuming an autoregressive process of order p.

Typically, large-sample asymptotic results are not helpful in small-sample situations. Going back to 1925, R. A. Fisher wrote, “Only by systematically tackling small sample problems on their merits does it seem possible to apply accurate tests to practical data” [9, 10]. This understanding led us to study the distribution and exact moments of LDA true error and comnon estimators [11, 3, 12, 13].

Having laid the groundwork for analyzing LDA related statistics in small-sample situations, in this work we establish a framework for studying LDA in stochastic settings, thereby allowing us to obtain the exact first and second moments of univariate LDA true error in a general stochastic 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, 7, 8] 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.

Another related line of research is the work on classification of stationary time series data [14, 15, 16]. The main focus in this work is to construct linear discriminant rules with the knowledge of having stationary data. In this framework the discriminant function is commonly the one which maximizes some measure of disparity between two multivariate densities, e.g. the Kullback-Leibler information measure. This means that the linear discriminant rules constructed here are no longer what is commonly known as LDA. Therefore, the main difference between the aforementioned results on studying the performance of LDA under correlated training data and the body of work on classification of stationary times series, is that the former focuses on the *analysis* of the effect of correlated training data (which may have a stationary structure) on the performance of LDA, and the latter focuses on the *synthesis* of new classification rules with the knowledge of having stationary time series. Our work is of the first type. We study the effect of training data that can be dependent and not necessarily identically distributed or stationary on the performance of LDA.

As an application of these results, we consider two commonly used models, first-order autoregressive and moving averages. We further study the exact effect of autoregressive or

moving-average model coefficients on changing the expected true error of LDA. Finally, we present numerical experiments to study several specific settings using the theory.

Before proceeding we note that univariate classification has played a major role in the history of pattern recogntion, in part, because of the ability to obtain closed-form solutions for error moments [1, 2, 3]; however, we should not overlook practical application. Indeed, most common tests for diagnosis and prognosis of cancer are univariate: PSA for prostate cancer [21], AFP for liver cancer [22], CA 125 for ovarian cancer [23], and CA 19.9 for colorectal cancer [24] are major protein markers. In addition to these protein biomarkers, there are genomic markers such as BRCA1 for breast cancer [25], BRCA2 [26] for male breast cancer, and APC for pancreatic cancer [27] that are major genomic markers.

2. Linear Discriminant Analysis and Error Estimation: Independent Sampling

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

$$W(\overline{X}^0, \overline{X}^1, X) = \frac{1}{\widehat{\sigma}^2}\left(X - \frac{\overline{X}^0 + \overline{X}^1}{2}\right)^T (\overline{X}^0 - \overline{X}^1), \quad (1)$$

where $\overline{X}^0 = \frac{1}{n_0}\sum_{i=1}^{n_0} X_i$ and $\overline{X}^1 = \frac{1}{n_1}\sum_{i=n_0+1}^{n_0+n_1} X_i$ are the sample means for each class and $\widehat{\sigma}^2$ is the pooled estimate of the variance of classes, which is assumed to be common in the LDA discriminant. Given X^0 and X^1, the designed LDA classifier is given by

$$\psi(X) = \begin{cases} 1, & \text{if } W(\overline{X}^0, \overline{X}^1, X) \le c \\ 0, & \text{if } W(\overline{X}^0, \overline{X}^1, X) > c \end{cases}, \quad (2)$$

with c being a constant. It is commonly assumed that c is zero [17], which is the assumption we also make throughout this paper. Therefore, the sign of W determines the classification of the sample point X and since $\widehat{\sigma}^2 > 0$, (1) reduces to

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

where $\overline{X} = \frac{\overline{X}^0 + \overline{X}^1}{2}$. Given the training data S_n (and thus X_0 and X_1), the classification error, also known as true error, is given by

$$\varepsilon = P(W(\overline{X}^0, \overline{X}^1, X) \le 0, X \in \Pi_0 | \overline{X}^0, \overline{X}^1) + P(W(\overline{X}^0, \overline{X}^1, X) > 0, X \in \Pi_1 | \overline{X}^0, \overline{X}^1) = \alpha_0 \varepsilon^0 + \alpha_1 \varepsilon^1, \quad (4)$$

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

$$\varepsilon^i = P((-1)^i W(\overline{X}^0, \overline{X}^1, X) \leq 0 | X \in \Pi_i, \overline{X}^0, \overline{X}^1). \quad (5)$$

The first and second moments of the classification error are given by

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

and

$$E[\varepsilon^2] = E[(\alpha_0 \varepsilon^0 + \alpha_1 \varepsilon^1)^2] = 2\alpha_0 \alpha_1 E[\varepsilon^0 \varepsilon^1] + \sum_{i=0}^{1} \alpha_i^2 E[\varepsilon^i \varepsilon^i]. \quad (7)$$

3. Performance of LDA classifier in Univariate Gaussian Dependent Sampling (UGDS) Model of Binary Classification

We now provide the mathematical framework to study LDA performance in a stochastic setting.

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 $[X_{t_1}, X_{t_2}, \ldots, X_{t_n}]^T$ has the multivariate normal distribution $N(\mu_T, \Sigma_T)$, where

$$\mu_T = [E(X_{t_1}), E(X_{t_2}), \ldots, E(X_{t_n})]^T = [\mu_1, \mu_2, \ldots, \mu_n]^T$$

and Σ_T is the covariance matrix dependent on $T = [t_1, t_2, \ldots, t_n]$.

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 $\mathbf{X}_{t^0}^0$ and $\mathbf{X}_{t^1}^1$ as $[X_{t_1^0}^0, X_{t_2^0}^0, \ldots, X_{t_{n_0}^0}^0, X_{t_1^1}^1, X_{t_2^1}^1, \ldots, X_{t_{n_1}^1}^1]^T$ possesses a multivariate normal distribution $N(\mu_T, \Sigma_T)$, where $\mu_T = [\mu_1^0, \mu_2^0, \ldots, \mu_{n_0}^0, \mu_1^1, \mu_2^1, \ldots, \mu_{n_1}^1]^T$, and

$$\sum\nolimits_T = \begin{bmatrix} \sum_{n_0 \times n_0}^{00} & \sum_{n_0 \times n_1}^{01} \\ \sum_{n_1 \times n_0}^{10} & \sum_{n_1 \times n_1}^{11} \end{bmatrix} \quad (8)$$

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^i. Thus, henceforth we denote $\mathbf{X}_{t_i}^i$ by $\mathbf{X}_t^i$ and the stacked vector $[X_{t_1^0}^0, X_{t_2^0}^0, \ldots, X_{t_{n_0}^0}^0, X_{t_1^1}^1, X_{t_2^1}^1, \ldots, X_{t_{n_1}^1}^1]^T$ by $[X_{t_1}^0, X_{t_2}^0, \ldots, X_{t_{n_0}}^0, X_{t_1}^1, X_{t_2}^1, \ldots, X_{t_{n_1}}^1]^T$.

Remark 1—If we assume $\mu_T = [\mu^{0T}, \mu^{1T}]^T$, with $\mu^i = [\mu^i, \mu^i, \ldots, \mu^i]_{1\times n_i}^T$, $\sum_{jj}^{ii} = (\sigma^i)^2$, $i = 0$, 1, $j = 1, 2, \ldots, n_i$, where $(\sigma^i)^2$ is the variance of class conditional distributions and $\sum_{jj}^{ii}$ indicates the diagonal elements of matrix $\mathbf{\Sigma}^{ii}$, $\sum_{jk}^{ii} = 0$, $i = 0$, 1, j, $k = 1, \ldots, n_i$, $j \neq k$, $\mathbf{\Sigma}^{01} = \mathbf{0}_{n_0\times n_1} = \mathbf{\Sigma}^{10T}$, and any future sample is independent from the training data and distributed either as $N(\mu^0, (\sigma^0)^2)$ or $N(\mu^1, (\sigma^1)^2)$, depending on its class, then the UGDS model reduces to the traditional i.i.d. sampling scenario defined in section 2. Because we will want to compare classifier errors in the dependent and independent scenarios, we will sometimes use ε^D and ε^I to denote errors in the respective settings.

Similar to (3), employing LDA with the UGDS model instead of traditional independent sampling in order to classify a sample point taken at t, denoted by X_t, results in the following W statistic for the univariate case

$$W(\overline{X}_T^0, \overline{X}_T^1, X_t) = (X_t - \overline{X}_T)(\overline{X}_T^0 - \overline{X}_T^1), \quad (9)$$

where $\overline{X}_T^0 = \frac{1}{n_0}\sum_{i=1}^{n_0} X_{t_i}^0$ and $\overline{X}_T^1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{t_i}^1$ are the sample means for each class and $\overline{X}_T = \frac{\overline{X}_T^0 + \overline{X}_T^1}{2}$. The designed LDA classifier is given by

$$\psi(X_t) = \begin{cases} 1, & \text{if } W(\overline{X}_T^0, \overline{X}_T^1, X_t) \leq 0 \\ 0, & \text{if } W(\overline{X}_T^0, \overline{X}_T^1, X_t) > 0 \end{cases}. \quad (10)$$

For the ease of notation, hereafter, we omit the subscript T from μ_T and $\mathbf{\Sigma}_T$.

3.1. Stochastic true error and its moments

Let $X_{t_s}^i$ denote a test sample point, 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

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

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 point, 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,

$$\varepsilon_{t_s} = \alpha_{t_s}^0 \varepsilon_{t_s}^0 + \alpha_{t_s}^1 \varepsilon_{t_s}^1, \quad (12)$$

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 $\varepsilon_{t_s}^i$ is the error rate specific to each process, with

$$\varepsilon_{t_s}^i = P((-1)^i W(\overline{X}_T^0, \overline{X}_T^1, X_{t_s}) \leq 0 | \overline{X}_T^0, \overline{X}_T^1, X_{t_s} \in \mathbf{X}_t^i). \quad (13)$$

By replacing $W(\overline{X}_T^0, \overline{X}_T^1, X_{t_s})$ 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:

$$E[\varepsilon_{t_s}] = \sum_{i=0}^{1} \alpha_{t_s}^i P((-1)^i W(\overline{X}_T^0, \overline{X}_T^1, X_{t_s}) \leq 0 | X_{t_s} \in \mathbf{X}_t^i). \quad (14)$$

In (12), the true error is indexed. One could, if desired, define the true error of a classifier to be the average error the classifier induces over an index set of interest, namely, $\varepsilon_{t_{s_1 - s_2}} = \frac{1}{s_2 - s_1} \sum_{s=s_1}^{s_2} \varepsilon_{t_s}$. Since characterizing ε_{t_s} yields a characterization of $\varepsilon_{t_{s_1 - s_2}}$, no generality is gained by averaging and we restrict our attention to ε_{t_s}. The second moment is also a function of t_s and from (12) we get

$$E[\varepsilon_{t_s}^2] = 2\alpha_{t_s}^0 \alpha_{t_s}^1 E[\varepsilon_{t_s}^0 \varepsilon_{t_s}^1] + \sum_{i=0}^{1} (\alpha_{t_s}^i)^2 E[(\varepsilon_{t_s}^i)^2]. \quad (15)$$

First focusing on $E[(\varepsilon_{t_s}^0)^2]$, the square of the probability defining $(\varepsilon_{t_s}^0)^2$ can be factored by introducing the random variable $X_{t_s}' \in \mathbf{X}_t^0$. Writing the probabilities as integrals of indicator functions allows us to apply Fubini's theorem, which shows X_{t_s} and X_{t_s}' to be independent (denoted $X_{t_s} \perp X_{t_s}'$). The expectation can then be applied. Altogether,

$$\begin{aligned} E[(\varepsilon_{t_s}^0)^2] &= E[P(W(\overline{X}_T^0, \overline{X}_T^1, X_{t_s}) \leq 0 | \overline{X}_T^0, \overline{X}_T^1, X_{t_s} \in \mathbf{X}_t^0)^2] \\ &= E\left[P(W(\overline{X}_T^0, \overline{X}_T^1, X_{t_s}) \leq 0 | \overline{X}_T^0, \overline{X}_T^1, X_{t_s} \in \mathbf{X}_t^0) \times P(W(\overline{X}_T^0, \overline{X}_T^1, X_{t_s}') \leq 0 | \overline{X}_T^0, \overline{X}_T^1, X_{t_s}' \in \mathbf{X}_t^0)\right] \\ &= E\left[P\left(W(\overline{X}_T^0, \overline{X}_T^1, X_{t_s}) \leq 0, W(\overline{X}_T^0, \overline{X}_T^1, X_{t_s}') \leq 0 | \overline{X}_T^0, \overline{X}_T^1, X_{t_s}, X_{t_s}' \in \mathbf{X}_t^0, X_{t_s} \perp X_{t_s}'\right)\right] \\ &= P\left(W(\overline{X}_T^0, \overline{X}_T^1, X_{t_s}) \leq 0, W(\overline{X}_T^0, \overline{X}_T^1, X_{t_s}') \leq 0 | X_{t_s}, X_{t_s}' \in \mathbf{X}_t^0, X_{t_s} \perp X_{t_s}'\right). \end{aligned} \quad (16)$$

$E[(\varepsilon_{t_s}^1)^2]$ and $E[\varepsilon_{t_s}^0 \varepsilon_{t_s}^1]$

can be expressed via similar factorizations. Hence,

$$\begin{aligned} E[\varepsilon_{t_s}^2] &= (\alpha_{t_s}^0)^2 P\left(W(\overline{X}_T^0, \overline{X}_T^1, X_{t_s}) \leq 0, W(\overline{X}_T^0, \overline{X}_T^1, X_{t_s}') \leq 0 | X_{t_s}, X_{t_s}' \in \mathbf{X}_t^0, X_{t_s} \perp X_{t_s}'\right) \\ &+ 2\alpha_{t_s}^0 \alpha_{t_s}^1 P\left(W(\overline{X}_T^0, \overline{X}_T^1, X_{t_s}) \leq 0, W(\overline{X}_T^0, \overline{X}_T^1, X_{t_s}') > 0 | X_{t_s} \in \mathbf{X}_t^0, X_{t_s}' \in \mathbf{X}_t^1, X_{t_s} \perp X_{t_s}'\right) \\ &+ (\alpha_{t_s}^1)^2 P\left(W(\overline{X}_T^0, \overline{X}_T^1, X_{t_s}) > 0, W(\overline{X}_T^0, \overline{X}_T^1, X_{t_s}') > 0 | X_{t_s}, X_{t_s}' \in \mathbf{X}_t^1, X_{t_s} \perp X_{t_s}'\right). \end{aligned} \quad (17)$$

To facilitate the subsequent discussion, we will explicitly denote the dependency of true error on the number of samples. Therefore, hereafter we use $\varepsilon_{t_s,n_0+n_1}$ and ε_{t_s}, or $\varepsilon^2_{t_s,n_0+n_1}$ and $\varepsilon^2_{t_s}$

, interchangeably.

Throughout the paper, we use the notations $Z < \mathbf{0}$ or $Z \geq \mathbf{0}$ to indicate componentwise inequalities, e.g. $Z = (z_1, z_2)^T < \mathbf{0}$ means $z_1 < 0$, $z_2 < 0$.

3.2. Expected performance of LDA in the UGDS model

The first moment of the classification error for LDA under the UGDS model is expressed exactly according to the following theorem.

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

$$E[\varepsilon^D_{t_s,n_0+n_1}] = \alpha^0_{t_s}\left[P(Z^1_s<\mathbf{0})+P(Z^1_s \geq \mathbf{0})\right] + \alpha^1_{t_s}\left[P(Z^{II}_s<\mathbf{0})+P(Z^{II}_s \geq \mathbf{0})\right], \quad (18)$$

where $Z^I_{t_s}$ and $Z^{II}_{t_s}$ are Gaussian bivariate vectors with

$$\mu_{Z^I_s} = [\ \mu^0_s - \tfrac{\bar{\mu}}{2} \quad -\mu'\ ]^T, \quad \mu_{Z^{II}_s} = [\ \mu^1_s - \tfrac{\bar{\mu}}{2} \quad \mu'\ ]^T,$$
$$\Sigma_{Z^I_s} = \begin{bmatrix} (\sigma^0_s)^2 - \frac{S_{\rho^{00}_s}}{n_0} - \frac{S_{\rho^{01}_s}}{n_1} + \frac{S_{\Sigma^{00}}}{4n_0^2} + \frac{S_{\Sigma^{11}}}{4n_1^2} + \frac{S_{\Sigma^{01}}}{2n_0n_1} & \frac{-S_{\rho^{00}_s}}{n_0} + \frac{S_{\rho^{01}_s}}{n_1} + \frac{S_{\Sigma^{00}}}{2n_0^2} - \frac{S_{\Sigma^{11}}}{2n_1^2} \\ \cdot & \frac{S_{\Sigma^{00}}}{n_0^2} + \frac{S_{\Sigma^{11}}}{n_1^2} - \frac{2S_{\Sigma^{01}}}{n_0n_1} \end{bmatrix},$$
$$\Sigma_{Z^{II}_s} = \begin{bmatrix} (\sigma^1_s)^2 - \frac{S_{\rho^{11}_s}}{n_1} - \frac{S_{\rho^{10}_s}}{n_0} + \frac{S_{\Sigma^{00}}}{4n_0^2} + \frac{S_{\Sigma^{11}}}{4n_1^2} + \frac{S_{\Sigma^{01}}}{2n_0n_1} & \frac{-S_{\rho^{11}_s}}{n_1} + \frac{S_{\rho^{10}_s}}{n_0} - \frac{S_{\Sigma^{00}}}{2n_0^2} + \frac{S_{\Sigma^{11}}}{2n_1^2} \\ \cdot & \frac{S_{\Sigma^{00}}}{n_0^2} + \frac{S_{\Sigma^{11}}}{n_1^2} - \frac{2S_{\Sigma^{01}}}{n_0n_1} \end{bmatrix}, \quad (19)$$

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

Proof: See the Appendix.

We note that under conditions stated in Remark 1, Theorem 1 reduces to Theorem 1 in [3].

Let $\Phi(x, y; \rho)$ be the cumulative bivariate normal distribution defined as:

$$\Phi(x, y; \rho) = \int_{-\infty}^{x}\int_{-\infty}^{y} \psi(u, v; \rho)\, du\, dv,$$
$$\psi(u, v; \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{\frac{-(u^2+v^2-2\rho uv)}{2(1-\rho^2)}\right\}. \quad (20)$$

We have the following Corollary.

Corollary 2—In the model considered in Theorem 1, let the training samples from each class have the same mean, that is $\mu = [\mu^{0^T}, \mu^{1^T}]^T$ in which $\mu^i = [\mu^i, \mu^i, \ldots, \mu^i]^T_{1\times n_i}$, $\mu^i_s = \mu^i$, $\sigma^i_s = \sigma$, $i = 0, 1$, meaning the test data at t_s has equal variances across classes, and $\alpha^0_{t_s} = \alpha^1_{t_s} = 0.5$. Furthermore, let $S_{\rho^{ik}_s} = 0$, $i, k = 0, 1$ Then

$$E[\varepsilon^{D}_{t_s,n_0+n_1}]=\frac{1}{2}-\frac{L(h,k;\rho)}{2}, \quad (21)$$

where

$$L(x,y;\rho)=\int_{-x}^{x}\int_{-y}^{y}\psi(u,v;\rho)du\,dv, \quad (22)$$

$$\begin{gathered} h=\frac{\mu}{2\sqrt{a}},\ k=\frac{\mu}{\sqrt{b}},\ \mu=\mu^0-\mu^1,\ \rho=\frac{\frac{S_{\Sigma^{00}}}{2n_0^2}-\frac{S_{\Sigma^{11}}}{2n_1^2}}{\sqrt{a}\sqrt{b}}, \\ a=\sigma^2+\frac{S_{\Sigma^{00}}}{4n_0^2}+\frac{S_{\Sigma^{11}}}{4n_1^2}+\frac{S_{\Sigma^{01}}}{2n_0n_1},\ b=\frac{S_{\Sigma^{00}}}{n_0^2}+\frac{S_{\Sigma^{11}}}{n_1^2}-\frac{2S_{\Sigma^{01}}}{n_0n_1}. \end{gathered} \quad (23)$$

Proof: See the Appendix.

To further proceed we present the following lemma, in which $\wedge$ denotes conjunction.

Lemma 3—Let $\Phi(x, y; \rho)$ be the cumulative bivariate normal distribution defined in (20) and defne $F(x, y; \rho)$ and $G(x, y; \rho)$ as follows:

$$\begin{aligned} F(x,y;\rho)&=\Phi(x,y;\rho)+\Phi(-x,-y;\rho), \\ G(x,y;\rho)&=F(x,y;\rho)+F(x,y;-\rho), \end{aligned} \quad (24)$$

where x and y are two constants such that $xy < 0$. Then, for $0 \le \lambda_x < 1$, $0 \le \lambda_y < 1$,

$$(|\rho_1| \le |\rho_0|) \wedge (x_1=\lambda_x x_0) \wedge (y_1=\lambda_y y_0) \Rightarrow G(x_1,y_1;\rho_1)>G(x_0,y_0;\rho_0). \quad (25)$$

Proof: See the Appendix.

Using this Lemma, we compare the expected true error of the UGDS model with the independent sampling model.

Corollary 4—In the model considered in Corollary 2, let $(\sigma^i_j)^2 \triangleq \sum^{ii}_{jj}$, $i = 0, 1$, $j = 1, 2, \ldots, n_i$, and let

$$\begin{aligned} \rho_D&=\frac{\frac{S_{\Sigma^{00}}}{n_0^2}-\frac{S_{\Sigma^{11}}}{n_1^2}}{\sqrt{(\sigma^2+\frac{S_{\Sigma^{00}}}{4n_0^2}+\frac{S_{\Sigma^{11}}}{4n_1^2}+\frac{S_{\Sigma^{01}}}{2n_0n_1})(\frac{S_{\Sigma^{00}}}{n_0^2}+\frac{S_{\Sigma^{11}}}{n_1^2}-\frac{2S_{\Sigma^{01}}}{n_0n_1})}}, \\ \rho_I&=\frac{\frac{\sum_{j=1}^{n_0}(\sigma^0_j)^2}{n_0^2}-\frac{\sum_{j=1}^{n_1}(\sigma^1_j)^2}{n_1^2}}{\sqrt{(\sigma^2+\frac{\sum_{j=1}^{n_0}(\sigma^0_j)^2}{4n_0^2}+\frac{\sum_{j=1}^{n_1}(\sigma^1_j)^2}{4n_1^2})(\frac{\sum_{j=1}^{n_0}(\sigma^0_j)^2}{n_0^2}+\frac{\sum_{j=1}^{n_1}(\sigma^1_j)^2}{n_1^2})}}. \end{aligned} \quad (26)$$

Let $E[\varepsilon^{I}_{t_s,n_0+n_1}]$ be the expectaton of the true error of the classifier in (10) specific to t_s and constructed as if all $n_0 + n_1$ training samples are i.i.d. (same mean and variance). Then

$$(|\rho_D| \leq |\rho_I|) \wedge \left(\frac{S'_{\Sigma^{00}}}{n_0^2} + \frac{S'_{\Sigma^{11}}}{n_1^2} \geq \max\{\frac{2S_{\Sigma^{01}}}{n_0 n_1}, \frac{-2S_{\Sigma^{01}}}{n_0 n_1}\} \right) \Rightarrow E[\varepsilon^D_{t_s,n_0+n_1}] \geq E[\varepsilon^I_{t_s,n_0+n_1}], \quad (27)$$

$$(|\rho_D| \geq |\rho_I|) \wedge \left(\frac{S'_{\Sigma^{00}}}{n_0^2} + \frac{S'_{\Sigma^{11}}}{n_1^2} \leq \min\{\frac{2S_{\Sigma^{01}}}{n_0 n_1}, \frac{-2S_{\Sigma^{01}}}{n_0 n_1}\} \right) \Rightarrow E[\varepsilon^D_{t_s,n_0+n_1}] \leq E[\varepsilon^I_{t_s,n_0+n_1}], \quad (28)$$

where S'_A is the sum of the off diagonal elements of matrix A, defined as $S'_A = \sum_{i,j,i\neq j} a_{ij}$

Proof: Find the expected true error in Theorem 1 using the conditions in the corollary and compare it to the expected true error determined by setting all off diagonal elements of $\mathbf{\Sigma}^{ii}$ to zero, $i = 0, 1$ and $\mathbf{\Sigma}^{01} = \mathbf{0}_{n_0 \times n_1}$. The proof follows by using the results of Lemma 3 in Theorem 1.

A more restricted set of sufficient conditions than those presented in Corollary 4 follows.

Corollary 5—In the model considered in Corollary 2, let $(\sigma^i_j)^2 \triangleq \sum^{ii}_{jj}$, $i = 0, 1$, $j = 1, 2, \ldots, n_i$, and

$$\frac{S_{\Sigma^{00}}}{n_0^2} - \frac{S_{\Sigma^{11}}}{n_1^2} = \frac{\sum_{j=1}^{n_0} (\sigma^0_j)^2}{n_0^2} - \frac{\sum_{j=1}^{n_1} (\sigma^1_j)^2}{n_1^2}. \quad (29)$$

Then

$$\frac{S'_{\Sigma^{00}}}{n_0^2} + \frac{S'_{\Sigma^{11}}}{n_1^2} \geq \max\{\frac{2S_{\Sigma^{01}}}{n_0 n_1}, \frac{-2S_{\Sigma^{01}}}{n_0 n_1}\} \Rightarrow E[\varepsilon^D_{t_s,n_0+n_1}] \geq E[\varepsilon^I_{t_s,n_0+n_1}], \quad (30)$$

$$\frac{S'_{\Sigma^{00}}}{n_0^2} + \frac{S'_{\Sigma^{11}}}{n_1^2} \geq \min\{\frac{2S_{\Sigma^{01}}}{n_0 n_1}, \frac{-2S_{\Sigma^{01}}}{n_0 n_1}\} \Rightarrow E[\varepsilon^D_{t_s,n_0+n_1}] \geq E[\varepsilon^I_{t_s,n_0+n_1}], \quad (31)$$

where S'_A is the sum of off diagonal elements of matrix A, defined as $S'_A = \sum_{i,j,i\neq j} a_{ij}$

Proof: The proof is similar to Corollary 4.

To have a sense of the conditions stated in Corollary 5, consider a scenario in which $n_0 = n_1$, the sample points in each class are equi-correlated with correlation ρ, and there is independent sampling across classes. This satisfies (29). If $\rho > 0$, then (30) holds and $E[\varepsilon^D_{t_s,n_0+n_1}] \geq E[\varepsilon^I_{t_s,n_0+n_1}]$. If $\rho < 0$ and the class covariance matrices are positive definite, then (31) holds and $E[\varepsilon^D_{t_s,n_0+n_1}] < E[\varepsilon^I_{t_s,n_0+n_1}]$.

Let us reflect on Corollaries 4 and 5. A correlated set of n sample points can be considered as a set in which the points convey some information about each other. Therefore, they are often considered to be as informative as n' independent samples with $n' < n$, thereby

producing a poorer classifier. This intuition aligns with the simple situation in which the sample points in each class are equi-correlated with $\rho > 0$ and the sample points across the two classes are uncorrelated. This scenario is a special case of (30) and $E[\varepsilon^D_{t_s,n_0+n_1}] \geq E[\varepsilon^I_{t_s,n_0+n_1}]$. However, (31) shows that there are correlation patterns that result in an expected true error smaller than it would be were there independent sampling, which means that sampling satisfying (31) is like having a larger sample size than if sampling were independent.

To illustrate, in the UGDS model suppose the training sample points are identically distributed as two Gaussian distribution, $N(-1, 1)$ for class 0 and $N(1, 1)$ for class 1. Let $n_0 = n_1 = 3$ and assume that any future test point is also distributed identically to the training data of its class. Furthermore, assume the data are generated via two different scenarios, *a* and *b*, such that $\mathbf{\Sigma}^{01} = \mathbf{0}_{3\times 3}$ and, for $i = 0, 1$,

$$\sum\nolimits_a^{ii} = \begin{bmatrix} 1 & -1/4 & -1/4 \\ -1/4 & 1 & \rho \\ -1/4 & \rho & 1 \end{bmatrix}, \quad \sum\nolimits_b^{ii} = \begin{bmatrix} 1 & 1/4 & 1/4 \\ 1/4 & 1 & \rho \\ 1/4 & \rho & 1 \end{bmatrix} \quad (32)$$

Figure 1(a) shows the expected true error of the classifier designed in scenario *a* as a function of ρ. It demonstrates that for some dependency patterns, as defined by the covariance matrix, the classifier has better performance than if the sampling were independent. Note that in Fig. 1(a) the curves meet at $\rho = 0.5$, the point of equality for the inequalities (30) and (31). Note also that for $\rho = -0.499$, $E[\varepsilon^D_6] = E[\varepsilon^I_{18}] = 0.165$. Hence, for the sampling covariance matrix (32), 3 points have the effect of 9 independent points. In general, better classification accuracy may be achieved if the sample points are collected according to specific schemes. Equations (28) and (31) provide sufficient sets of conditions that result in such schemes.

Figure 1(b) shows the expected true error of a classifier constructed in scenario *b* by varying ρ in the same range as in scenario *a*. The only difference between scenarios *a* and *b* is changing the covariances between the first sample point and other sample points to positive values. It results in the curve for dependent sampling in Figure 1(b) being substantially above the curve for independent sampling.

3.3. Second moment of LDA true error in the UGDS model

Next we obtain the second moment of true error of LDA at t_s as defined in (17).

Theorem 6—Under the UGDS model, the second moment of LDA at t_s is

$$E[(\varepsilon^D_{t_s,n_0+n_1})^2] = (\alpha^0_{t_s})^2 \left[P(Z^I_s < \mathbf{0}) + P(Z^I_s \geq \mathbf{0})\right] + 2\alpha^0_{t_s}\alpha^I_{t_s} \left[P(Z^{II}_s < \mathbf{0}) + P(Z^{II}_s \geq \mathbf{0})\right] + (\alpha^1_{t_s})^2 \left[P(Z^{III}_s < \mathbf{0}) + P(Z^{III}_s \geq \mathbf{0})\right], \quad (33)$$

where Z^j_s is a 3-variate Gaussian random vector with mean and covarianc matrices as follows:

$$\mu_{Z^I_s} = [\ \mu^0_s - \tfrac{\bar{\mu}}{2} \quad -\mu' \quad \mu^0_s - \tfrac{\bar{\mu}}{2}\ ]^T, \mu_{Z^{II}_s} = [\ \mu^0_s - \tfrac{\bar{\mu}}{2} \quad -\mu' \quad -\mu^1_s + \tfrac{\bar{\mu}}{2}\ ]^T \quad (34)$$

and, for $i, j = 0, 1$, $i \neq j$, letting

$$z_s^i=(\sigma_s^i)^2-\frac{S_{\rho_s^{ii}}}{n_i}-\frac{S_{\rho_s^{ij}}}{n_j}+\frac{S_{\Sigma^{00}}}{4n_0^2}+\frac{S_{\Sigma^{11}}}{4n_1^2}+\frac{S_{\Sigma^{01}}}{2n_0n_1}, \quad (35)$$

we have

$$\sum_{Z_s^I}=\begin{bmatrix} z_s^0 & \frac{-S_{\rho_s^{00}}}{n_0}+\frac{S_{\rho_s^{01}}}{n_1}+\frac{S_{\Sigma^{00}}}{2n_0^2}-\frac{S_{\Sigma^{11}}}{2n_1^2} & z_s^0-(\sigma_s^0)^2 \\ \cdot & \frac{S_{\Sigma^{00}}}{n_0^2}+\frac{S_{\Sigma^{11}}}{n_1^2}-\frac{2S_{\Sigma^{01}}}{n_0n_1} & \frac{-S_{\rho_s^{00}}}{n_0}+\frac{S_{\rho_s^{01}}}{n_1}+\frac{S_{\Sigma^{00}}}{2n_0^2}-\frac{S_{\Sigma^{11}}}{2n_1^2} \\ \cdot & \cdot & z_s^0 \end{bmatrix},$$
$$\sum_{Z_s^{II}}=\begin{bmatrix} z_s^0 & \frac{-S_{\rho_s^{00}}}{n_0}+\frac{S_{\rho_s^{01}}}{n_1}+\frac{S_{\Sigma^{00}}}{2n_0^2}-\frac{S_{\Sigma^{11}}}{2n_1^2} & -z_s^0+(\sigma_s^0)^2+\frac{S_{\rho_s^{11}}-S_{\rho_s^{01}}}{2n_1}+\frac{S_{\rho_s^{10}}-S_{\rho_s^{00}}}{2n_0} \\ \cdot & \frac{S_{\Sigma^{00}}}{n_0^2}+\frac{S_{\Sigma^{11}}}{n_1^2}-\frac{2S_{\Sigma^{01}}}{n_0n_1} & \frac{-S_{\rho_s^{11}}}{n_1}+\frac{S_{\rho_s^{10}}}{n_0}-\frac{S_{\Sigma^{00}}}{2n_0^2}+\frac{S_{\Sigma^{11}}}{2n_1^2} \\ \cdot & \cdot & z_s^1 \end{bmatrix}, \quad (36)$$

with $\bar{\mu}=\frac{\sum_{i=1}^{n_0}\mu_i^0}{n_0}+\frac{\sum_{i=1}^{n_1}\mu_i^1}{n_1}$, $\mu'=\frac{\sum_{i=1}^{n_0}\mu_i^0}{n_0}-\frac{\sum_{i=1}^{n_1}\mu_i^1}{n_1}$, S_A is the sum of all elements of matrix A, defined as $S_A=\Sigma_{i,j}a_{ij}$, and μ_s^i and $(\sigma_s^i)^2$ are the mean and variance of random variables at t_s from class i, $i=0, 1$. Furthermore, $\mu_{Z^{III}}$ is obtained from μ_{Z^I} by replacing $-\mu'$ with μ' and μ_s^0 with μ_s^1, and $\sum_{Z_s^{III}}$ is obtained from $\sum_{Z_s^I}$ by exchanging n_0 and n_1, $(\sigma_s^0)^2$ and $(\sigma_s^1)^2$, $S_{\boldsymbol{\Sigma}^{00}}$ and $S_{\boldsymbol{\Sigma}^{11}}$, $S_{\rho_s^{00}}$ and $S_{\rho_s^{11}}$, and $S_{\rho_s^{01}}$ and $S_{\rho_s^{10}}$.

Proof: See the Appendix.

Let $E[\varepsilon_{t_s,n_0+n_1}^I]$ and $E[(\varepsilon_{t_s,n_0+n_1}^I)^2]$ be the first and second moments of true error of the classifier in (10) specific to t_s and constructed by n_0+n_1 independent training sample points distributed according to the same mean and variance. Then we have the following corollary.

Corollary 7—In the model considered in corollary 5, further assume $n_0=n_1=n$, $\boldsymbol{\Sigma}^{01}=\mathbf{0}_{n\times n}$, and, for $k, j=1, 2, \ldots, n$,

$$\sum_{kj}^{00}=\sum_{kj}^{11}=\begin{cases}\sigma^2 & k=j\\ \rho>0 & otherwise\end{cases}, \quad (37)$$

where σ is the common variance of test sample points across classes at t_s. Let m be the number of additional dependent training points in each class with the same class conditional means and dependency structure, meaning $\sum_{kj}^{ii}$ as in (37) for $k, j=1, 2, \ldots, n+m$ and $\boldsymbol{\Sigma}^{01}=\mathbf{0}_{(n+m)\times(n+m)}$, that are required to make $E[\varepsilon_{t_s,2n+2m}^D]=E[\varepsilon_{t_s,2n}^I]$. This number also makes $E[(\varepsilon_{t_s,2n+2m}^D)^2]=E[(\varepsilon_{t_s,2n}^I)^2]$ and is given by

$$m=\frac{n-1}{\frac{\sigma^2}{n\rho}-1} \quad (38)$$

Proof: The proof of $E[\varepsilon_{t_s,2n+2m}^D]=E[\varepsilon_{t_s,2n}^I]$ follows by equating elements of covariance matrices obtained for the dependent model in (19) with the covariance matrices for the

independent sampling model. Under the conditions of the corollary, these matrices in the independent sampling scenario (given by Theorem 1 in [3]) are

$$\sum_{Z_s^I} = \sum_{Z_s^{II}} = \sum_{Z_s^{III}} = \begin{pmatrix} \sigma^2 + \frac{\sigma^2}{2n} & 0 \\ 0 & \frac{2\sigma^2}{n} \end{pmatrix} \quad (39)$$

Furthermore, we note that the conditions stated in this corollary satisfy the condition stated in (30), and hence $E[\varepsilon^D_{t_s,2n}] \geq E[\varepsilon^I_{t_s,2n}]$. The proof of $E[(\varepsilon^D_{t_s,2n+2m})^2] = E[(\varepsilon^I_{t_s,2n})^2]$ follows similarly by equating covariance matrices presented in (36) with those presented in Theorem 2 in [3].

In (38), if $\rho > \frac{\sigma^2}{n}$, then $m < 0$, meaning that adding any additional points under the dependency model in the corollary does not lower $E[\varepsilon^D_{t_s,2n+2m}]$ and $E[(\varepsilon^D_{t_s,2n+2m})^2]$ to the level of the first and second moments of true error of the constructed LDA classifier as if the original $2n$ training samples were independent.

4. Applications

In this section we study applications to common models used in signal processing, the first-order autoregressive model, AR(1), and the first-order moving average model, MA(1), by assuming the training data are generated by the output processes of two models. $\mathbf{Z}_t^0$ and $\mathbf{Z}_t^1$ are two independent *white noise* processes and $\mathbf{X}_t^0$ and $\mathbf{X}_t^1$ are the processes producing the system output. The goal is to characterize the performance of the LDA classifier as a function of sample size, the parameters of the white noise processes, and the autoregressive coefficients.

4.1. First-order autoregressive model AR(1)

We consider two AR(1) models:

$$X_t^i = c_i + \psi_i X_{t-1}^i + Z_t^i, \quad i = 0, 1, \quad (40)$$

where ψ_i is a constant such that $0 < |\psi_i| < 1$, $i = 0, 1$, and $Z_t^0 \sim N(0, \sigma_0^2)$, $Z_t^1 \sim N(0, \sigma_1^2)$, for all t, 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 8—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 (40). Then, at t_s, where $\max\{n_0, n_1\} < s$, the expected true error of LDA constructed using the training samples $[X_{t_1}^0, X_{t_2}^0, \ldots, X_{t_{n_0}}^0]$ and $[X_{t_1}^1, X_{t_2}^1, \ldots, X_{t_{n_1}}^1]$ is

$$E[\varepsilon^{AR(1)}_{t_s,n_0+n_1}] = \alpha^0_{t_s}\left[P(Z_s^I < \mathbf{0}) + P(Z_s^I \geq \mathbf{0})\right] + \alpha^I_{t_s}\left[P(Z_s^{II} < \mathbf{0}) + P(Z_s^{II} \geq \mathbf{0})\right], \quad (41)$$

where $Z^I_{t_s}$ and $Z^{II}_{t_s}$ are Gaussian bivariate vectors with

$$\mu_{Z_s^I}=[\ \tfrac{\mu}{2}\quad -\mu\ ]^T,\quad \mu_{Z_s^{II}}=[\ \tfrac{-\mu}{2}\quad \mu\ ]^T,$$
$$\sum_{Z_s^I}=\begin{bmatrix}\frac{\sigma_0^2}{1-\psi_0^2}-\frac{S_{\rho_s}^{00}}{n_0}+\frac{S_{\sum}^{00}}{4n_0^2}+\frac{S_{\sum}^{11}}{4n_1^2} & \frac{-S_{\rho_s}^{00}}{n_0}+\frac{S_{\sum}^{00}}{2n_0^2}-\frac{S_{\sum}^{11}}{2n_1^2}\\ \cdot & \frac{S_{\sum}^{00}}{n_0^2}+\frac{S_{\sum}^{11}}{n_1^2}\end{bmatrix},\quad (42)$$
$$\sum_{Z_s^{II}}=\begin{bmatrix}\frac{\sigma_1^2}{1-\psi_1^2}-\frac{S_{\rho_s}^{11}}{n_1}+\frac{S_{\sum}^{00}}{4n_0^2}+\frac{S_{\sum}^{11}}{4n_1^2} & \frac{-S_{\rho_s}^{11}}{n_1}-\frac{S_{\sum}^{00}}{2n_0^2}+\frac{S_{\sum}^{11}}{2n_1^2}\\ \cdot & \frac{S_{\sum}^{00}}{n_0^2}+\frac{S_{\sum}^{11}}{n_1^2}\end{bmatrix},$$

where

$$\mu=\frac{c_0}{1-\psi_0}-\frac{c_1}{1-\psi_1},\ S_{\rho_s}^{ii}=\frac{\psi_i^{(s-n_i)}\sigma_i^2}{1-\psi_i^2}\left(\frac{1-\psi_i^{n_i}}{1-\psi_i}\right),$$
$$S_{\sum}^{ii}=\frac{\sigma_i^2}{(1-\psi_i^2)(1-\psi_i)}\left[n_i(1+\psi_i)-2\psi_i\left(\frac{1-\psi_i^{n_i}}{1-\psi_i}\right)\right]. \quad (43)$$

Proof: See the Appendix.

Corollary 9—In the model considered in Theorem 8, let $\psi_0 = \psi_1 = \psi$, $\sigma_0 = \sigma_1$, $\alpha_{t_s}^0=\alpha_{t_s}^1$. Let $E[\varepsilon_{t_s,n_0+n_1}^{AR(1)_\psi}]$ denote the expected true error of an AR(1) model with AR coefficient ψ specific to t_s. Then

$$\lim_{s\to\infty}E[\varepsilon_{t_s,n_0+n_1}^{AR(1)_\psi}]=\frac{1}{2}-\frac{L(h,k;\rho)}{2}, \quad (44)$$

where $L(h, k; \rho)$ is defined in (22) and

$$h=\frac{\mu}{2\sqrt{a}},\ k=\frac{\mu}{\sqrt{b}},\ \mu=\frac{c_0-c_1}{1-\psi},\ a=\frac{\sigma^2}{1-\psi^2}+\frac{b}{4},$$
$$b=\frac{\sigma^2\left[(1+\psi)\left(\frac{1}{n_0}+\frac{1}{n_1}\right)-2\frac{\psi}{1-\psi}\left(\frac{1-\psi^{n_0}}{n_0^2}+\frac{1-\psi^{n_1}}{n_1^2}\right)\right]}{(1-\psi^2)(1-\psi)}, \quad (45)$$
$$\rho=\frac{\sigma^2\left[(1+\psi)\left(\frac{1}{n_0}-\frac{1}{n_1}\right)-\frac{2\psi}{1-\psi}\left(\frac{1-\psi^{n_0}}{n_0^2}-\frac{1-\psi^{n_1}}{n_1^2}\right)\right]}{2(1-\psi^2)(1-\psi)\sqrt{a}\sqrt{b}}.$$

Proof: See the Appendix.

We consider $E[\varepsilon_{t_s,2n}^{AR(1)}]$ as a function of ψ and compare it to the case where ψ= 0, which corresponds to the stochastic i.i.d setting.

Corollary 10—In the model considered in Corollary 9, let $n_0 = n_1 = n$. Furthermore, let $E[\varepsilon_{t_s,2n}^{I}]$ be the expected true error of the LDA classifier with ψ = 0 in (40). Let ψ' and ψ'' be two arbitrary values of the AR coefficient ψ. Then

$$\psi''>\psi' \Rightarrow \lim_{s\to\infty}E[\varepsilon_{t_s,2n}^{AR(1)_{\psi''}}]<\lim_{s\to\infty}E[\varepsilon_{t_s,2n}^{AR(1)_{\psi'}}]. \quad (46)$$

Hence,

$$0<\psi<1 \Rightarrow \lim_{s\to\infty} E[\varepsilon^{AR(1)_\psi}_{t_s,2n}] < \lim_{s\to\infty} E[\varepsilon^{I}_{t_s,2n}],$$
$$-1<\psi<0 \Rightarrow \lim_{s\to\infty} E[\varepsilon^{AR(1)_\psi}_{t_s,2n}] > \lim_{s\to\infty} E[\varepsilon^{I}_{t_s,2n}]. \quad (47)$$

Proof: See the Appendix.

Corollary 10 shows that, if $\psi \in (0, 1)$, then under the conditions of the Corollary, constructing an LDA classifier to differentiate between AR processes is beneficial in terms of the expected true error tested on sufficiently lagged data; however, for $\psi \in (-1, 0)$, we expect larger expected true error.

4.2. First-order moving-average model MA(1)

We consider the MA(1) models

$$X^i_t = c_i + Z^i_t + \theta_i Z^i_{t-1}, \quad i=0, 1, \quad (48)$$

where $\theta_i \in \mathbb{R}$ and $Z^0_t \sim N(0, \sigma^2_0), Z^1_t \sim N(0, \sigma^2_1)$, for all t, are independent from each other. Then $\mathbf{X}^0_t = \{X^0_t : 0<t<\infty\}$ and $\mathbf{X}^1_t = \{X^1_t : 0<t<\infty\}$

are two independent and covariance-stationary processes regardless of the values of θ_i [18].

Theorem 11—Let $\mathbf{X}^0_t, \mathbf{X}^1_t$ in the UGDS model be defined by the two independent covariance-stationary MA(1) processes defined in (48). Then, at t_s, where $\max\{n_0, n_1\} + 1 < s$, the expected true error of an LDA classifier constructed using the training samples $[X^0_{t_1}, X^0_{t_2}, \ldots, X^0_{t_{n_0}}]$ and $[X^1_{t_1}, X^1_{t_2}, \ldots, X^1_{t_{n_1}}]$ is

$$E[\varepsilon^{MA(1)}_{t_s,n_0+n_1}] = \alpha^0_{t_s}\left[P(Z^I_s<\mathbf{0}) + P(Z^I_s \geq \mathbf{0})\right] + \alpha^1_{t_s}\left[P(Z^{II}_s<\mathbf{0}) + P(Z^{II}_s \geq \mathbf{0})\right], \quad (49)$$

where $Z^I_{t_s}$ and $Z^{II}_{t_s}$ are Gaussian bivariate vectors with:

$$\mu_{Z^I_s} = \begin{bmatrix} \frac{\mu}{2} & -\mu \end{bmatrix}^T, \quad \mu_{Z^{II}_s} = \begin{bmatrix} \frac{-\mu}{2} & \mu \end{bmatrix}^T,$$
$$\sum_{Z^I_s} = \begin{bmatrix} \sigma^2_0(1+\theta^2_0) + \frac{S_{\sum^{00}}}{4n^2_0} + \frac{S_{\sum^{11}}}{4n^2_1} & \frac{S_{\sum^{00}}}{2n^2_0} - \frac{S_{\sum^{11}}}{2n^2_1} \\ \cdot & \frac{S_{\sum^{00}}}{n^2_0} + \frac{S_{\sum^{11}}}{n^2_1} \end{bmatrix},$$
$$\sum_{Z^{II}_s} = \begin{bmatrix} \sigma^2_1(1+\theta^2_1) + \frac{S_{\sum^{00}}}{4n^2_0} + \frac{S_{\sum^{11}}}{4n^2_1} & \frac{-S_{\sum^{00}}}{2n^2_0} - \frac{S_{\sum^{11}}}{2n^2_1} \\ \cdot & \frac{S_{\sum^{00}}}{n^2_0} + \frac{S_{\sum^{11}}}{n^2_1} \end{bmatrix}, \quad (50)$$

where $i = 0, 1$, and

$$\mu = c_0 - c_1, \; S_{\sum^{ii}} = \sigma^2_i\left[n_i(1+\theta^2_i) + 2(n_i-1)\theta_i\right]. \quad (51)$$

Proof: See the Appendix.

Corollary 12—For the model in Theorem 11, let $\theta_0 = \theta_1$, $\sigma_0 = \sigma_1$, $\alpha_{t_s}^0 = \alpha_{t_s}^1$. Let $E[\varepsilon_{t_s,n_0+n_1}^{MA(1)_\theta}]$ denote the expected true error of an MA(1) model with MA coefficient θ specific to t_s. Then

$$E[\varepsilon_{t_s,n_0+n_1}^{MA(1)_\theta}] = \frac{1}{2} - \frac{L(h,k;\rho)}{2}, \quad (52)$$

where $L(h, k; \rho)$ is defined in (22) and

$$\begin{gathered} h = \frac{\mu}{2\sqrt{a}},\ k = \frac{\mu}{\sqrt{b}},\ \mu = c_0 - c_1,\ a = \sigma^2(1+\theta^2) + \frac{b}{4}, \\ b = \sigma^2\left[(1+\theta)^2\left(\frac{1}{n_0}+\frac{1}{n_1}\right) - 2\theta\left(\frac{1}{n_0^2}+\frac{1}{n_1^2}\right)\right], \\ \rho = \frac{\sigma^2\left[(1+\theta)^2\left(\frac{1}{n_0}-\frac{1}{n_1}\right) - 2\theta\left(\frac{1}{n_0^2}-\frac{1}{n_1^2}\right)\right]}{2\sqrt{a}\sqrt{b}}. \end{gathered} \quad (53)$$

Proof: The result follows by considering the assumption of the corollary in Theorem 11 and then following the same steps similar to Corollary 2.

Corollary 13—For the model in Corollary 12, let $n_0 = n_1 = n$. Furthermore, let $E[\varepsilon_{t_s,2n}^I]$ be the expected true error of the LDA classifier specific to t_s with $\theta = 0$ in (48). Let θ and θ' be two arbitrary values of the MA coefficient θ. Then

$$\begin{gathered} (\theta'' > \theta') \wedge \left(\theta', \theta'' \in (-\infty, \tfrac{1}{n} - 1)\right) \Rightarrow E[\varepsilon_{t_s,2n}^{MA(1)_{\theta''}}] < E[\varepsilon_{t_s,2n}^{MA(1)_{\theta'}}], \\ (\theta'' > \theta') \wedge \left(\theta', \theta'' \in [\tfrac{\frac{1}{n}-1}{2n+1}, \infty)\right) \Rightarrow E[\varepsilon_{t_s,2n}^{MA(1)_{\theta''}}] > E[\varepsilon_{t_s,2n}^{MA(1)_{\theta'}}], \end{gathered} \quad (54)$$

and, therefore,

$$\begin{gathered} \theta \in (0, \infty) \Rightarrow E[\varepsilon_{t_s,2n}^{MA(1)_\theta}] > E[\varepsilon_{t_s,2n}^I], \\ \theta \in [\tfrac{\frac{1}{n}-1}{2n+1}, 0) \Rightarrow E[\varepsilon_{t_s,2n}^{MA(1)_\theta}] < E[\varepsilon_{t_s,2n}^I]. \end{gathered} \quad (55)$$

Proof: See the Appendix.

Corollary 13 shows that there exists a range of moving-average coefficients, i.e. $[\frac{\frac{1}{n}-1}{2n+1}, 0)$, that is beneficial in terms of expected classification error, i.e. has a smaller expected true error than the stochastic i.i.d model. For positive values of the coefficient, the expected true error of LDA increases.

5. Numerical Examples

We now illustrate the results obtained in previous sections under several specific settings.

Experiment 1

First, we consider scenarios in which the sample points taken from each class conditional process are identically distributed. They have the same mean, μ_0 for class 0 and μ_1 for class 1, and we set $\mu_0 = -\mu_1$ and $\mu_0 = 0.5, 0.75, 1, 1.5$. We assume that the observations have variance 1 and are equally correlated with $\rho_{with} \in [\rho_l, 0.95]$. The value of ρ_l is determined so

that the covariance matrix defined in (8) is positive definite. In each case we consider three settings for the correlation, ρ_{bet}, across classes: (1) independent, $\rho_{bet} = 0$, (2) $\rho_{bet} = \rho_{with}$, and (3) $\rho_{bet} = -\rho_{with}$. For each setting we consider two sample sizes, $n_0 = n_1 = n = 5$ and $n_0 = n_1 = n = 25$. We assume any future observation from each class conditional process has a distribution similar to those of the training data from that class and $\alpha^0_{t_s} = \alpha^1_{t_s}$.

Figure 2(a)–2(d) show the exact expectation and standard deviation (SD) of the LDA true error for this experiment as a function of ρ_{with}. The results are calculated from Theorems 1 and 6. Parts a and b of the figure show that increasing ρ_{with} has an incremental effect on $E[\varepsilon^D_{t_s,2n}]$. Since future observations are identically distributed, $E[\varepsilon^D_{t_s,2n}]$ is the same for all values of t_s. Theoretically, for $\rho_{bet} = 0$, we can easily verify the graphical behavior by using Lemma 3 in Theorem 1. To analytically see the effect of ρ_{with} on $E[\varepsilon^D_{t_s,2n}]$ once $\rho_{bet} = 0$, let ρ_1, ρ_2 be two arbitrary values of ρ_{with} such that $\rho_1 < \rho_2$ and denote all distributional parameters used in Corollary 2 corresponding to ρ_k, $k = 1, 2$, with a super script ρ_k. With the aforementioned conditions of the experiment, we have $\frac{s^{\rho_1}_{\Sigma^{00}}}{n_0^2} - \frac{s^{\rho_1}_{\Sigma^{11}}}{n_1^2} = \frac{s^{\rho_2}_{\Sigma^{00}}}{n_0^2} - \frac{s^{\rho_2}_{\Sigma^{11}}}{n_1^2} = 0$ and $\frac{s^{\rho_k}_{\Sigma^{00}}}{n_0^2} + \frac{s^{\rho_k}_{\Sigma^{11}}}{n_1^2} = \frac{1+2(n-1)\rho_k}{2n}$, $k = 1, 2$. Therefore, $a^{\rho 1} < a^{\rho 2}$ and $b^{\rho 1} < b^{\rho 2}$. The results then follow from Lemma 3. For other cases where $\rho_{bet} \neq 0$ one may analytically study the effect of changing ρ_{bet} on $E[\varepsilon^D_{t_s,2n}]$ using results Theorem 1 and studying the change similar to the proof of Lemma 3.

Figures 2(a) and 2(b) show that increasing $d = |\mu_0 - \mu_1|$ has an incremental effect on $E[\varepsilon^D_{t_s,2n}]$. This effect can also be seen from Lemma 3 and Corollary 2. Therefore, we call classification scenarios with a larger d, "easier" scenarios, and those with smaller d, "harder" scenarios. In this sense, d is an indicator of classification difficulty in our experiment. The figures suggest that having a between-class correlation of $\rho_{bet} = \rho_{with} > 0$ helps in classification performance in "harder" classification situations (i.e., compared with $\rho_{bet} = 0$) and has a detrimental effect on classification performance in "easier" settings. However, having $\rho_{bet} = \rho_{with} < 0$, helps to have a better classification performance in "easier" settings and results in a worse performance in "harder" settings. This is observed by the fact that curves for $\rho_{bet} = \rho_{with}$ are above (below) the curves for $\rho_{bet} = 0$ for $d = 3$ ($d = 1$).

The standard deviation is more complicated to interpret. The trends seen in Figures 2(c) and 2(d) suggest that increasing ρ_{with} generally increases the standard deviation of the LDA true error in cases where $\rho_{bet} = 0$. Furthermore, it suggests that once $\rho_{bet} = -\rho_{with}$, the standard deviation generally increases as ρ_{with} grows, but once $\rho_{bet} = \rho_{with}$, increasing ρ_{with} in small sample sizes may increase or decrease the standard deviation depending on classification difficulty, and as the sample size gets larger, increasing ρ_{with} generally increases the standard deviation. Furthermore, the figures suggest that increasing the classification difficulty may first increase the standard deviation and then decreases it.

Comparing Figure 2(a) with 2(b) and Figure 2(c) with 2(d) shows that increasing sample sizes lower the magnitude of the expectation and standard deviation regardless of classification difficulty or magnitude of ρ_{with}.

Experiment 2

In this experiment, we use the first order autoregressive model defined in (40). 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 2(e) and 2(f) show the exact expectation of

LDA true error for this experiment. These results are exact and are calculated from Theorem 8. These figures suggest that increasing ψ decreases $E[\varepsilon^{AR(1)_\psi}_{t_s,2n}]$. According to Corollary 10, for a sufficiently lagged t_s, $E[\varepsilon^{AR(1)_\psi}_{t_s,2n}]$ is a decreasing function of ψ and, furthermore, $E[\varepsilon^{AR(1)_\psi}_{t_s,2n}]<E[\varepsilon^{I}_{t_s,2n}]$ for $0 < \psi < 1$ and $E[\varepsilon^{AR(1)_\psi}_{t_s,2n}]>E[\varepsilon^{I}_{t_s,2n}]$ for $0 < \psi < 1$. Here the same behavior is observed even for small lags of 2 and 10. Furthermore, decreasing the sample size and increasing the classification difficulty have an incremental effect on the expected true error.

Experiment 3

In this experiment, we use the AR(1) model in (48). We assume $\alpha^0_{t_s}=\alpha^1_{t_s}$, $n_0 = n_1 = n$, $\sigma_0 = \sigma_1 = 1$, and $\theta_0 = \theta_1 = \theta \in [-10, 10]$. We consider $c_0 = 0.5, 0.75, 1, 1.5$ with $c_0 = -c_1$. Figure 3 shows the exact expectation of the LDA true error for this experiment. These results are exact and are calculated from Theorem 11.

Figure 3 shows that the expected true error of LDA under the MA(1) model has an inverted bell shape with a negatively biased center, and the bias decreases as the sample size increases. The results of Corollary 13 are clear in this figure: for $\theta \in (-\infty, \frac{1}{n}-1]$, $E[\varepsilon^{MA(1)_\theta}_{t_s,2n}]$ is a decreasing function of θ. This region is on the left-hand side of the vertical blue dotted lines in Figure 3. For $\theta \in [\frac{\frac{1}{n}-1}{2n+1}\infty)$, $E[\varepsilon^{MA(1)_\theta}_{t_s,2n}]$ is an increasing function of θ. This region is on the right-hand side of the vertical red dashed line in the figure. As proved in Corollary 13, we observe in Figure 3(c) and 3(d) that, for $\theta \in [\frac{\frac{1}{n}-1}{2n+1}, 0)$, $E[\varepsilon^{MA(1)_\theta}_{t_s,2n}]<E[\varepsilon^{I}_{t_s,2n}]$. This is the region between red dashed line and the vertical black line of each plot. For $\theta \in (0, \infty)$, $E[\varepsilon^{MA(1)_\theta}_{t_s,2n}]>E[\varepsilon^{I}_{t_s,2n}]$. This is the region on the right-hand side of the vertical black solid lines.

Experiment 4

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 [19]. Following [20], we divide the 307 individuals used in this study into 64 “poor” prognosis (class 0) versus 243 “good” prognosis (class 1) patients. A poor prognosis is defined to be a distant metastasis within 5 years of initial diagnosis. 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. Using this dataset 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. We assume the single variate used in this classifier is the ALDH4 gene, which has the highest correlation with prognosis of breast cancer in [20]. Therefore, in this scenario, the experimenter is given 12 “technical” replicates in total, which are now treated as our “sample points”. This is an example of the UGDS model in genomic applications in which our classification is defined by two Gaussian processes, $\mathbf{X}^0_t$ and $\mathbf{X}^1_t$, which are assumed to be independent processes. We note that the expected performance of a classifier depends on t_s, i.e. $E[\varepsilon^D_{t_s,12}]$ in Theorem 1, which is a function of the distribution of the future data as well as the distribution of the training data and their correlation structure. We verify the Gaussianity of each of the 12 random variables, $X^0_{t_1}, X^0_{t_2}, \ldots, X^0_{t_6}, X^1_{t_1}, X^1_{t_2}, \ldots, X^1_{t_6}$, used for

characterizing the two Gaussian processes of this example via a Shapiro-Wilk test (using the R statistical software) on the full dataset corresponding to each random variable. This test does 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.

Unfortunately, taken together, the 12 random variables do not pass the Shapiro-Wilk test for multivariate Gaussianity. Nonetheless, we will proceed and demonstrate that, even with this lack of multivariate Gaussianity, Theorem 1 is much more accurate than its counterpart in [3], which assumes i.i.d. data from each distribution.

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, $E[\varepsilon^{D}_{t_s,12}]$, to differentiate samples distributed as $X^0_{t_5}$ from samples distributed as $X^1_{t_1}$ is 0.475. To further verify this expected performance we construct a classifier on each possible combination among 243 × 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. It is interesting to compare this accuracy to the case in which one designs a classifier without paying attention to the correlation structure between samples and various distributions governing the data (considering the data being i.i.d.). In this scenario one (incorrectly) considers the data from each class coming from a single distribution and the expected performance of a classifier can be therefore evaluated from Theorem 1 that we presented in [3]. Again we use the sample means and variances, computed on the full dataset, as estimates of the unknown true means and variances. In this case the expected performance of LDA is estimated to be 0.374, which is very far from 0.479.

6. Conclusion

In many applications, the assumption of having i.i.d. training samples is violated. This paper characterizes the performance of univariate LDA classification in stochastic settings by assuming the samples are taken from two class conditional Gaussian processes, which are not necessarily independent. Linear classification has been considered owing to its long history in pattern recognition and its suitability for small-sample classification. We do not impose a specific correlation structure on the training data. We have presented conditions in which the correlation structure can be either beneficial or detrimental in terms of classification performance. As an application we have obtained exact expressions for the performance of LDA in situations that the data are produced through auto-regressive (AR) or moving-average (MA) models of the first order. We have found ranges of AR or MA multiplicative coefficients having incremental or decremental effect on classification performance. Having characterized univariate LDA performance in closed form, we aim to follow our work in [3] and characterize the effect of non-i.i.d. samples on training-data-based error estimators.

Acknowledgments

This work was partially supported by the NIH grants 2R25CA090301 (Nutrition, Biostatistics, and Bioinformatics) from the National Cancer Institute.

References

1. Hills M. Allocation rules and their error rates. J Royal Statist Soc Ser B (Methodological). 1966; 28(1):1–31.
2. Sorum MJ. Estimating the expected probability of misclassification for a rule based on the linear discriminant function: Univariate normal case. Technometrics. 1973; 15:329–339.
3. Zollanvari A, Braga-Neto UM, Dougherty ER. 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. 2012; 45:908–917.
4. Basu JP, Odell PL. Effect of intraclass correlation among training samples on the misclassification probabilities of bayes' procedure. Pattern Recogn. 1974; 6:13–16.
5. McLachlan GJ. Further results on the effect of interclass correlation among training samples in discriminant analysis. Pattern Recogn. 1976; 8:273–275.
6. Tubbs JD. Effect of autocorrelated training samples on bayes' probability of mis-classlficatlon. Pattern Recogn. 1980; 12:351–354.
7. Lawoko CRO, McLachlan GJ. Discrimination with autocorrelated observations. Pattern Recogn. 1985; 18:145–149.
8. Lawoko CRO, McLachlan GJ. Asymptotic error rates of the w and z statistics when the training observations are dependent. Pattern Recogn. 1986; 19:467–471.
9. Fisher, RA. Statistical Methods for Research Workers. Edinburgh: Oliver &Boyd; 1925.
10. Martin JK, Hirschberg DS. Small sample statistics for classification error rates ii: Confidence intervals and significance tests. 1996
11. Zollanvari A, Braga-Neto UM, Dougherty ER. On the sampling distribution of resubstitution and leave-one-out error estimators for linear classifiers. Pattern Recogn. 2009; 42(11):2705–2723.
12. Zollanvari A, Braga-Neto UM, Dougherty ER. Joint sampling distribution between actual and estimated classification errors for linear discriminant analysis. IEEE Trans Inf Theory. 2010; 56(2):784–804.
13. Zollanvari A, Braga-Neto UM, Dougherty ER. Analytic study of performance of error estimators for linear discriminant analysis. IEEE Trans Sig Proc. 2011; 59(9):4238–4255.
14. Shumway RH, Unger AN. Linear discriminant functions for stationary time series. J Am Statist Assoc. 1974; 69:948–956.
15. Kakizawa Y, Shumway R, Taniguchi M. Discrimination and clustering for multivariate time series. J Am Statist Assoc. 1998; 93:328–340.
16. Kazakos D, Papantoni-Kazakos P. Spectral distance measuring between gaussian processes. IEEE Trans Autom Control. 1980; 25:950–959.
17. McLachlan GJ. The asymptotic distributions of the conditional error rate and risk in discriminant analysis. Biometrika. 1974; 61:131–135.
18. Hamilton, JD. Time Series Analysis. NJ: Princeton University Press; 1994.
19. Buyse M, Loi S, 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. 2006; 98:1183–1192. [PubMed: 16954471]
20. vanÕt, Veer L.; Dai, H., et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002; 415:530–6. [PubMed: 11823860]
21. Schršder FH, Hugosson J, et al. Screening and prostate-cancer mortality in a randomized european study. New Eng J Med. 2009; 360:1320–1328. [PubMed: 19297566]
22. Koelinka CJL, van Hasseltb P, et al. Tyrosinemia type i treated by ntbc: How does afp predict liver cancer? Mol Genet Metab. 2006; 89:310–315. [PubMed: 17008115]
23. Bast RC, Xu FJ, et al. Ca 125: the past and the future. Int J Biol Markers. 1998; 13:179–187. [PubMed: 10228898]
24. Filella X, Molina R, et al. Prognostic value of ca 19.9 levels in colorectal cancer. Mol Genet Metab. 1992; 216:55–59.

25. Frank TS, Deffenbaugh AM, et al. Clinical characteristics of individuals with germline mutations in brca1 and brca2: Analysis of 10,000 individuals. J Clin Oncol. 2002; 20:1480–1490. [PubMed: 11896095]
26. Syrjakoski K, Kuukasjarvi T, et al. Brca2 mutations in 154 finnish male breast cancer patients. Neoplasia. 2004; 6:541–545. [PubMed: 15548363]
27. Horii A, Nakatsuru S, et al. Frequent somatic mutations of the apc gene in human pancreatic cancer. Cancer Res. 1992; 52:6696–6698. [PubMed: 1423316]
28. Abramowitz, M.; Stegun, IA. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. New York: Dover Publications; 1972.
29. Scheffe H. A useful convergence theorem for probability distributions. Ann Math Statist. 1947; 18:434–438.
30. Anderson B, Jackson J, Sitharam M. A useful convergence theorem for probability distributions. Amer Math Monthly. 1998; 105:447–451.

Appendix. Various Correlation Structures

Let p-dimenional sample points of each class, $X_1, X_2, \ldots, X_{n_i}$, with X_j being a column vector, be separately taken from two p-variate normal distributions, Π_1 and Π_2, with the distribution $N(\mu_i, \mathbf{\Sigma})$. Furthermore, let V_i be the dispersion matrix of the $n_i p \times 1$ vector $X=(X_1^T, X_2^T, \ldots, X_{n_i}^T)^T$, $i = 0, 1$, defined as $V_i = E[(X - E(X))(X - E(X))^T]$. We define three correlation structures in regard to the data: (1) equicorrelated if $V_i = I_{n_i} \otimes (\mathbf{\Sigma} - R) + E_{n_i} \otimes R$, with R being a symmetric matrix, I_n the $n \times n$ identity matrix, and E_n the $n \times n$ matrix with all elements being 1; (2) simply equicorrelated if $V_i = I_{n_i} \otimes (1 - \rho)\mathbf{\Sigma} + E_{n_i} \otimes \rho\mathbf{\Sigma}$, where ρ is a nonzero scalar constant where $|\rho| < 1$; and (3) serially correlated if $V_i = I_{n_i} \otimes \mathbf{\Sigma} + E_{n_i} \otimes \rho_\tau \mathbf{\Sigma}$, where $\tau = |k - l|$, $k, l = 1, 2, \ldots, n_i$, $|\rho_\tau| < 1$, $\tau = 1, 2, \ldots, n_i$, $\rho_0 = 0$. Note that univariate sample points, equicorrelation and simple-equicorrelation structures are essentially the same.

Proof of Theorem 1

From (9), it follows that

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

where $\overline{X}_T = \frac{\overline{X}_T^0 + \overline{X}_T^1}{2}$. Expanding $\overline{X}_T^0$ and $\overline{X}_T^1$ 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

$$E[\varepsilon_{t_s}^0] = P(Z_s^I < \mathbf{0}) + P(Z_s^I \geq \mathbf{0}), \quad (56)$$

where $Z_s^I = A Y_s^0$ 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

$$A = \begin{bmatrix} 1 & -\frac{1}{2n_0} & \frac{-1}{2n_0} & \cdots & \frac{-1}{2n_0} & \frac{-1}{2n_1} & \cdots & \frac{-1}{2n_1} \\ 0 & \frac{-1}{n_0} & \frac{-1}{n_0} & \cdots & \frac{-1}{n_0} & \frac{1}{n_1} & \cdots & \frac{1}{n_1} \end{bmatrix}. \quad (57)$$

Therefore, Z_s^I is a Gaussian random vector with mean $A\mu_{\gamma_s^0}$ and covariance matrix $A\sum_{\gamma_s^0}A^T$. Plugging in the values of $\mu_{\gamma_s^0}=[\mu_s^0,\mu_1^0,\mu_2^0,\ldots,\mu_{n_0}^0,\mu_1^1,\ldots,\mu_{n_1}^1]$ and noting the fact that the j^{th} element of vector ρ_s^{ik} is defined as $\rho_s^{ik}(j)=E[(X_{t_s}^i-\mu_s^i)(X_{t_j}^k-\mu_j^k)]$, $i, k = 0, 1, j = 1, 2, \ldots, n_k$, we have

$$\sum_{\gamma_s^0}=\begin{bmatrix}(\sigma_s^0)^2 & \rho_s^{00} & \rho_s^{01}\\ (\rho_s^{00})^T & \sum^{00} & \sum^{01}\\ (\rho_s^{01})^T & (\sum^{01})^T & \sum^{11}\end{bmatrix}, \quad (58)$$

which leads to the expression stated in Theorem 1. Evaluating the mean and covariance matrix of vector Z_s^{II}, which is the counterpart for $E[\varepsilon_{t_s}^1]$, is entirely similar, by considering $P(W(\overline{X}_T^0,\overline{X}_T^1,X_{t_s})>0|\overline{X}_T^0,\overline{X}_T^1,X_{t_s}\in\mathbf{X}_t^1)$.

Proof of Corollary 2

Note that for $\Phi(x, y; \rho)$ defined in (20),

$$\Phi(-x,-y;\rho)=\int_x^\infty\int_y^\infty\frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left\{\frac{-(u^2+v^2-2\rho uv)}{2(1-\rho^2)}\right\}dudv. \quad (59)$$

By considering the assumption of the corollary for Theorem 1, and using (20) and (59) in (18), we get

$$E[\varepsilon_{t_s,n_0+n_1}^D]=\frac{1}{2}\left(\Phi(\frac{\mu}{2\sqrt{a}},\frac{-\mu}{\sqrt{b}},\rho)+\Phi(\frac{-\mu}{2\sqrt{a}},\frac{\mu}{\sqrt{b}},\rho)\right)+\frac{1}{2}\left(\Phi(\frac{-\mu}{2\sqrt{a}},\frac{\mu}{\sqrt{b}},-\rho)+\Phi(\frac{\mu}{2\sqrt{a}},\frac{-\mu}{\sqrt{b}},-\rho)\right), \quad (60)$$

with a, b, and ρ defined in the corollary. Using the identity [28]

$$2[\Phi(x,y;\rho)+\Phi(x,y;-\rho)-\Phi(x)-\Phi(y)]+1=L(x,y;\rho), \quad (61)$$

where $\Phi(.)$ is the standard normal cumulative function, completes the proof.

Proof of Lemma 3

Here, we first provide a way to intuitively understand the Lemma and then we provide a rigorous proof. We have

$$G(x,y;\rho)=F(x,y;\rho)+F(x,y;-\rho)=1+L(x,y,\rho)=1-L(|x|,|y|,\rho), \quad (62)$$

where the last equality is due to $xy < 0$ stated as an assumption to the lemma. Intuitively, the lemma makes sense because smaller values of $|x|$, $|y|$, and $|\rho|$ imply not only a smaller integration region in (22), but also less mass in that region. Next we provide a rigorous proof. It is straightforward to show

$$\begin{aligned}\frac{\partial G(x,y;\rho)}{\partial x}&=\frac{2e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\left[\Phi\left(\frac{y-\rho x}{\sqrt{1-\rho^2}}\right)+\Phi\left(\frac{y+\rho x}{\sqrt{1-\rho^2}}\right)-1\right],\\ \frac{\partial G(x,y;\rho)}{\partial y}&=\frac{2e^{-\frac{y^2}{2}}}{\sqrt{2\pi}}\left[\Phi\left(\frac{x-\rho y}{\sqrt{1-\rho^2}}\right)+\Phi\left(\frac{x+\rho y}{\sqrt{1-\rho^2}}\right)-1\right],\\ \frac{\partial G(x,y;\rho)}{\partial \rho}&=2\psi(x,y;\rho)-2\psi(x,y;-\rho),\end{aligned} \tag{63}$$

where the last equality comes from well known results of Gaussian distribution, where $\frac{\partial\Phi(x,y;\rho)}{\partial\rho}=\frac{\Phi(x,y;\rho)}{\partial x\partial y}=\psi(x,y;\rho)$. Without loss of generality, we assume $x\geq 0$ and $y\leq 0$. The results for $x\leq 0$ and $y\geq 0$ are entirely similar after exchanging x and y in the following proof. We have

$$\begin{aligned}&\rho\geq 0\Rightarrow y-\rho x\leq 0, y-\rho x\leq y+\rho x\leq -y+\rho x\\ &\Rightarrow \Phi\left(\frac{y-\rho x}{\sqrt{1-\rho^2}}\right)+\Phi\left(\frac{y+\rho x}{\sqrt{1-\rho^2}}\right)\leq 1\Rightarrow \frac{\partial G}{\partial x}\leq 0,\\ &\rho<0\Rightarrow y+\rho x\leq 0, y+\rho x\leq y-\rho x\leq -y-\rho x\\ &\Rightarrow \Phi\left(\frac{y-\rho x}{\sqrt{1-\rho^2}}\right)+\Phi\left(\frac{y+\rho x}{\sqrt{1-\rho^2}}\right)\leq 1\Rightarrow \frac{\partial G}{\partial x}\leq 0.\end{aligned} \tag{64}$$

Hence, $\frac{\partial G}{\partial x}\leq 0$. Similarly, $\frac{\partial G}{\partial y}\geq 0$. Furthermore,

$$\rho\geq 0\Rightarrow\frac{\partial G(x,y;\rho)}{\partial\rho}\leq 0,\quad \rho<0\Rightarrow\frac{\partial G(x,y;\rho)}{\partial\rho}>0.$$

For $0\leq\lambda\leq 1$, we set

$$\gamma_x=\lambda x_1+(1-\lambda)x_0,\ \gamma_y=\lambda y_1+(1-\lambda)y_0,\ \gamma_\rho=\lambda\rho_1+(1-\lambda)\rho_0. \tag{65}$$

Then λ, $\gamma_x\geq 0$, $\gamma_y\leq 0$, $\rho_i\geq 0\Rightarrow\gamma_\rho\geq 0$ ($i=0, 1$), and $\rho_i<0\Rightarrow\gamma_\rho<0$ ($i=0, 1$). Thus, $\frac{\partial G}{\partial\gamma_x}\leq 0$, $\frac{\partial G}{\partial\gamma_y}\geq 0$ and

$$\gamma_\rho\geq 0\Rightarrow\frac{\partial G(x,y;\rho)}{\partial\gamma_\rho}\leq 0,\quad \gamma_\rho<0\Rightarrow\frac{\partial G(x,y;\rho)}{\partial\gamma_\rho}>0. \tag{66}$$

Then

$$\begin{aligned}\frac{dG}{d\lambda}&=\frac{\partial G}{\partial\gamma_x}\frac{d\gamma_x}{d\lambda}+\frac{\partial G}{\partial\gamma_y}\frac{d\gamma_y}{d\lambda}+\frac{\partial G}{\partial\gamma_\rho}\frac{d\gamma_\rho}{d\lambda}\\ &=\frac{\partial G}{\partial\gamma_x}(x_1-x_0)+\frac{\partial G}{\partial\gamma_y}(y_1-y_0)+\frac{\partial G}{\partial\gamma_\rho}(\rho_1-\rho_0).\end{aligned} \tag{67}$$

First assume $\rho_i\geq 0$, $i=0, 1$, so that $\gamma_\rho\geq 0$, $\frac{\partial G(x,y;\rho)}{\partial\gamma_\rho}\leq 0$. Since $\frac{\partial G}{\partial\gamma_x}\leq 0$, $x_1\leq x_0$, $\frac{\partial G}{\partial\gamma_y}\geq 0$, $y_1\geq y_0$, and $\rho_1\leq\rho_0$, we have $\frac{dG}{d\lambda}\geq 0$. Therefore,

$$\int_0^1\frac{dG}{d\lambda}d\lambda=G(1)-G(0)=G(x_1,y_1,\rho_1)-G(x_0,y_0,\rho_0)\geq 0. \tag{68}$$

Next assume $\rho_i\leq 0$, $i=0, 1$, so that $\gamma_\rho\leq 0$, $\frac{\partial G(x,y;\rho)}{\partial\gamma_\rho}\geq 0$. Since $\frac{\partial G}{\partial\gamma_x}\leq 0$, $x_1\leq x_0$, $\frac{\partial G}{\partial\gamma_y}\geq 0$, $y_1\geq y_0$, and $\rho_1\geq\rho_0$, we have $\frac{dG}{d\lambda}\geq 0$. Therefore,

$$\int_0^1 \frac{dG}{d\lambda} d\lambda = G(1) - G(0) = G(x_1, y_1, \rho_1) - G(x_0, y_0, \rho_0) \geq 0.$$

Lastly, assume the ρ_i's have opposite signs. Without loss of generality, assume $\rho_0 < 0$, $\rho_1 \geq 0$, and $|\rho_1| \leq |\rho_0|$. Then

$$\begin{aligned} \int_0^1 \frac{dG}{d\lambda} d\lambda &= \int_0^{\frac{|\rho_0|-\rho_1}{\rho_1-\rho_0}} dG + \int_{\frac{|\rho_0|-\rho_1}{\rho_1-\rho_0}}^1 dG \\ &= \int_0^{\frac{|\rho_0|-\rho_1}{\rho_1-\rho_0}} dG + G(x_m, y_m, -\rho_1) - G(x_1, y_1, \rho_1), \end{aligned} \tag{69}$$

where $x_m = \frac{|\rho_0|-\rho_1}{\rho_1-\rho_0} x_1 + (1 - \frac{|\rho_0|-\rho_1}{\rho_1-\rho_0}) x_0$, $y_m = \frac{|\rho_0|-\rho_1}{\rho_1-\rho_0} y_1 + (1 - \frac{|\rho_0|-\rho_1}{\rho_1-\rho_0}) y_0$, $x_1 < x_m < x_0$, and $y_0 < y_m < y_1$. From the definition of $G(x, y, \rho)$ it is easy to see that $G(x, y, \rho) = G(x, y, -\rho)$ and then, from the conditions that result in (68), we have $G(x_m, y_m, -\rho_1) - G(x_1, y_1, \rho_1) = G(x_m, y_m, \rho_1) - G(x_1, y_1, \rho_1) \geq 0$. Hence, in order to show $G(x_1, y_1, \rho_1) - G(x_0, y_0, \rho_0) \geq 0$ it is sufficient to show that $\int_0^{\frac{|\rho_0|-\rho_1}{\rho_1-\rho_0}} \frac{dG}{d\lambda} d\lambda \geq 0$. For $\lambda \in [0, \frac{|\rho_0|-\rho_1}{\rho_1-\rho_0}]$, we have $\gamma_\rho \leq 0$, $\frac{\partial G(x,y;\rho)}{\partial \gamma_\rho} \geq 0$. Therefore, from (67), $\frac{\partial G(x,y;\rho)}{\partial \gamma_\rho} \geq 0$ and, furthermore, $\frac{dG}{d\lambda} \geq 0$. Thus, $\int_0^{\frac{|\rho_0|-\rho_1}{\rho_1-\rho_0}} \frac{dG}{d\lambda} d\lambda \geq 0$ and the result follows.

Proof of Theorem 6

From (16) and (9), it follows that

$$E[(\varepsilon_{t_s}^0)^2] = P(X_{t_s} - \overline{X}_T < 0, \overline{X}_T^0 - \overline{X}_T^1 > 0, X_{t_s}' - \overline{X}_T < 0) + P(X_{t_s} - \overline{X}_T \geq 0, \overline{X}_T^0 - \overline{X}_T^1 < 0, X_{t_s}' - \overline{X}_T \geq 0), \tag{70}$$

where $\overline{X}_T = \frac{\overline{X}_T^0 + \overline{X}_T^1}{2}$. Expanding $\overline{X}_T^0$ and $\overline{X}_T^1$ 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

$$E[(\varepsilon_{t_s}^0)^2] = P(Z_s^I < \mathbf{0}) + P(Z_s^I \geq \mathbf{0}), \tag{71}$$

where $Z_s^I = A Y_s^0$, in which $Y_s^0 = [X_{t_s}^0, X_{t_s}'^0, X_{t_1}^0, \ldots, X_{t_{n_0}}^0, X_{t_1}^1, \ldots, X_{t_{n_1}}^1]^T$, and the super index 0 in $X_{t_s}^0$ and $X_{t_s}'^0$ is to denote explicitly $X_{t_s}, X_{t_s}' \in \mathbf{X}_t^0$, and

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

Therefore, Z_s^I is a Gaussian random vector with mean $A\mu_{Y_s^0}$ and covariance matrix $A \sum_{Y_s^0} A^T$. Plugging in the values of $\mu_{Y_s^0} = [\mu_s^0, \mu_s^0, \mu_1^0, \mu_2^0, \ldots, \mu_{n_0}^0, \mu_1^1, \mu_2^1, \ldots, \mu_{n_1}^1]$ and noting the fact that the j^{th} element of vector $\rho_s^{ik}(j)$ is defined as $\rho_s^{ik}(j) = E[(X_{t_s}^i - \mu_s^i)(X_{t_j}^k - \mu_j^k)]$, $i, k = 0, 1$, $j = 1, 2, \ldots, n_k$, and from the definition of $X_{t_s}'^0$ it holds that $E[(X_{t_s}^0 - \mu_s^i)(X_{t_j}^k - \mu_j^k)] = E[(X_{t_s}'^0 - \mu_s^i)(X_{t_j}^k - \mu_j^k)]$, we have:

$$\sum_{Y_s^0} = \begin{bmatrix} (\sigma_s^0)^2 & 0 & \rho_s^{00} & \rho_s^{01} \\ 0 & (\sigma_s^0)^2 & \rho_s^{00} & \rho_s^{01} \\ (\rho_s^{00})^T & (\rho_s^{00})^T & \sum^{00} & \sum^{01} \\ (\rho_s^{01})^T & (\rho_s^{01})^T & (\sum^{01})^T & \sum^{11} \end{bmatrix}, \quad (72)$$

which leads to the expression stated in Theorem 6. Evaluating the mean and covariance matrices of vector Z_s^{II} and Z_s^{III} is entirely similar.

Proof of Theorem 8

Since the Z_t^i's are Gaussian, X_t^0 and X_t^1 are covariance-stationary [18] and the vectors $X_{n_0}^0=[X_{t_1}^0, X_{t_2}^0, \ldots, X_{t_{n_0}}]^T$ and $X_{n_0}^0=[X_{t_1}^1, X_{t_2}^1, \ldots, X_{t_{n_1}}^1]^T$ are distributed normally as $X_{n_i}^i \sim N(\mu^i, \sum^i)$, $i = 0, 1$, where

$$\begin{gathered} \mu^i=[\mu^i, \mu^i, \ldots, \mu^i]_{1\times n_i}^T, \ \sum^i(k,l)=\frac{\psi_i^{|k-l|}}{1-\psi_i^2}\sigma_i^2, \ k,l=1,2,\ldots,n_i, \\ \rho_s^{ii}(k)=\frac{\psi_i^{s-k}}{1-\psi_i^2}\sigma_i^2, \ k=1,2,\ldots,n_i, \ \rho_s^{01}=\mathbf{0}_{1\times n_1}, \ \rho_s^{10}=\mathbf{0}_{1\times n_0}, \end{gathered} \quad (73)$$

$\mu_i=\frac{c_i}{1-\psi_i}$, and $\Sigma^i(k, l)$ denotes the entry in the k^{th} row and l^{th} column of matrix Σ^i. The result follows by replacing (73) in Theorem 1.

Proof of Corollary 9

Using the corollary assumptions in Theorem 8, we get

$$E[\varepsilon_{t_s,n_0+n_1}^{AR(1)_\psi}]=\frac{1}{2}\sum_{i=0}^{1}\left(\Phi(h_s^i,-k;\rho_s^i)+\Phi(-h_s^i,k;\rho_s^i)\right), \quad (74)$$

$$\begin{gathered} h_s^i=\frac{\mu}{2\sqrt{a_s^i}}, \quad \rho_s^i=-\frac{\psi^{(s-n_i)}\sigma^2}{n_i(1-\psi^2)}\left(\frac{1-\psi^{n_i}}{1-\psi}\right)+\rho, \\ a_s^i=a-\frac{\psi^{(s-n_i)}\sigma^2}{n_i(1-\psi^2)}\left(\frac{1-\psi^{n_i}}{1-\psi}\right), \end{gathered} \quad (75)$$

with k, a, b, ρ, and μ defined in (45). Let $F(x, y; \rho) = \Phi(x, y; \rho) + \Phi(-x, -y; \rho)$, with $\Phi(x, y; \rho)$ defined in (20). Then using Scheffe's Lemma [29] we have

$$2\lim_{s\to\infty}E[\varepsilon_{t_s,n_0+n_1}^{AR(1)_\psi}]=F(\lim_{s\to\infty}h_s^0,-k;\lim_{s\to\infty}\rho_s^0)+F(\lim_{s\to\infty}h_s^1,-k;\lim_{s\to\infty}\rho_s^1). \quad (76)$$

Note that by taking the limit, the term $\frac{\psi^{(s-n_i)}\sigma^2}{n_i(1-\psi^2)}\left(\frac{1-\psi^{n_i}}{1-\psi}\right)$ in h_s^i and a_s^i converges exponentially to 0 and we have $a_s^0=a_s^1=a$, $h_s^0=h_s^1=h$, $\rho_s^0=\rho_s^1=\rho$. The result follows similarly to proof of Corollary 2.

Proof of Corollary 10

With $n_0 = n_1 = n$, we have $\rho = 0$ in (45). From (76) we get $2\lim_{s\to\infty} E[\varepsilon_{t_s,2n}^{AR(1)_\psi}] = G(h_\psi, l_\psi; 0)$, with $G(h_\psi, -k_\psi; 0)$ defined as in (24) and $l_\psi = -k_\psi$, where we use a subscript ψ to explicitly denote dependence of l and h on ψ. Since $h_\psi l_\psi < 0$, we can use a proof similar to that of Lemma 3 to compare different AR models. Specifically, suppose we prove that

$$\psi'' > \psi' \Rightarrow \exists \lambda_h \in [0, 1) \wedge \exists \lambda_l \in [0, 1): h_{\psi''} = \lambda_h h_{\psi'} \wedge l_{\psi''} = \lambda_l l_{\psi'} \quad (77)$$

Then, similar to proof of Lemma 3, we can prove $G(h_{\psi''}, l_{\psi''}; 0) < G(h_{\psi'}, l_{\psi'}; 0)$, so that

$$2\lim_{s\to\infty} E[\varepsilon_{t_s,2n}^{AR(1)_{\psi''}}] = G(h_{\psi''}, l_{\psi''}; 0) < \lim_{s\to\infty} E[\varepsilon_{t_s,2n}^{AR(1)_{\psi'}}] = G(h_{\psi'}, l_{\psi'}; 0), \quad (78)$$

thereby proving the basic inequality in the corrollary. We first demonstrate (77). Assume $c_0 > c_1$. We first prove that for $\psi \in (-1, 1)$, we have $\frac{dl_\psi}{d\psi} < 0$ and $\frac{dh_\psi}{d\psi} > 0$. It is easy to see that:

$$\frac{dl_\psi}{d\psi} = -\sqrt{\frac{n}{2}}\frac{(c_0 - c_1)}{\sigma}\frac{d(\sqrt{f_\psi})}{d\psi} = -\sqrt{\frac{n}{2f_\psi}}\frac{(c_0 - c_1)g_\psi}{\sigma(d_\psi)^2},$$

where

$$g_\psi = 2n(1 + \psi^2 - (n+1)\psi^n + (n-1)\psi^{(n+2)}),$$
$$d_\psi = n - 2\psi - n\psi^2 + 2\psi^{n+1}, \quad f_\psi = \frac{n - n\psi^2}{d_\psi}. \quad (79)$$

From Descartes' Rule of Signs [30], g_ψ has either zero or two positive roots. For $n \geq 2$,

$$g_\psi = 4n(\psi - 1)^2\left(\frac{(n-1)}{2}\psi^n + \sum_{j=1}^{n-1} j\psi^j + \frac{1}{2}\right). \quad (80)$$

Therefore, for all n, g_ψ has two roots at 1 and these are the only positive roots. Similarly we observe that if n is even, then g_ψ has only two negative roots at −1. If n is odd, again from Descartes' Rule of Signs [30], g_ψ has only one negative root, denoted by ψ_-. We show that $\psi_- \in (-\infty, -1)$. Let $\psi_- = -\psi_+$, $\psi_+ > 0$. Since n is odd, we need to have

$$1 + \psi_+^2 + (n+1)\psi_+^n - (n-1)\psi_+^{(n+2)} = 0. \quad (81)$$

Were $\psi_+ \in (0, 1)$, this would imply $(n+1)\psi_+^n > (n-1)\psi_+^{(n+2)}$. Hence, (81) is not possible and $\psi_- \in (-\infty, -1)$. Summarizing this result, we see that $\psi \in (-1, 1) \Rightarrow g_\psi > 0$ and therefore, $\frac{dl_\psi}{d\psi} < 0$. It is straightforward to show

$$\frac{dh_\psi}{d\psi} = \sqrt{\frac{n}{2}}\frac{(c_0 - c_1)}{\sigma}\frac{d(\sqrt{r_\psi})}{d\psi} = \sqrt{\frac{n}{2r_\psi}}\frac{(c_0 - c_1)(4n^3(1-\psi)^2 + g_\psi)}{\sigma(2n^2(1-\psi)^2 + d_\psi)^2},$$

where $r_\psi = \frac{n-n\psi^2}{2n^2(1-\psi)^2+d_\psi}$. Since for $\psi \in (-1, 1)$ we have $g_\psi > 0$, then $\frac{dh_\psi}{d\psi} > 0$. We set $\gamma_\psi = \lambda\psi_1 + (1 - \lambda)\psi_0$, where $0 \le \lambda \le 1$. Now we check that (78) holds. Denoting $G(h_{\gamma_\psi}, l_{\gamma_\psi}; 0)$ by G, we have

$$\begin{aligned}\frac{dG}{d\lambda} &= \frac{\partial G}{\partial h_{\gamma_\psi}}\frac{dh_{\gamma_\psi}}{d\gamma_\psi}\frac{d\gamma_\psi}{d\lambda} + \frac{\partial G}{\partial l_{\gamma_\psi}}\frac{dl_{\gamma_\psi}}{d\gamma_\psi}\frac{d\gamma_\psi}{d\lambda} \\ &= \frac{\partial G}{\partial h_{\gamma_\psi}}\frac{dh_{\gamma_\psi}}{d\gamma_\psi}(\psi'' - \psi') + \frac{\partial G}{\partial l_{\gamma_\psi}}\frac{dl_{\gamma_\psi}}{d\gamma_\psi}(\psi'' - \psi').\end{aligned} \quad (82)$$

Since $\psi \in (-1, 1)$, $0 \le \lambda \le 1$, and $h_\psi l_\psi < 0$, we can see that $\gamma_\psi \in (-1, 1)$, $h_{\gamma_\psi} l_{\gamma_\psi} < 0$, $\frac{dh_{\gamma_\psi}}{d\gamma_\psi} > 0$, $\frac{dl_{\gamma_\psi}}{d\gamma_\psi} < 0$, and from Proof of Lemma 3 in the appendix, $\frac{\partial G}{\partial h_{\gamma_\psi}} \le 0$ and $\frac{\partial G}{\partial l_{\gamma_\psi}} \ge 0$. Since $\psi'' > \psi'$, we see that $\frac{dG}{d\lambda} < 0$. Similar to the proof of Lemma 3, integrating over λ results in $G(h_{\psi''}, l_{\psi''}; 0) < G(h_{\psi'}, l_{\psi'}; 0)$. The same basic argument goes through for $c_0 < c_1$ and we have $\frac{dh_{\gamma_\psi}}{d\gamma_\psi} < 0$, $\frac{dl_{\gamma_\psi}}{d\gamma_\psi} > 0$. The remaining results follow from the definition of $E[\varepsilon^I_{t_s,2n}]$, where we have $E[\varepsilon^{AR(1)_{\psi=0}}_{t_s,2n}] = E[\varepsilon^I_{t_s,2n}]$.

Proof of Theorem 11

Since the Z_t^i's are Gaussian, X_t^0 and X_t^1 are covariance-stationary and the vectors $X^0_{n_0} = [X^0_{t_1}, X^0_{t_2}, \ldots, X_{t_{n_0}}]^T$ and $X^0_{n_0} = [X^1_{t_1}, X^1_{t_2}, \ldots, X^1_{t_{n_1}}]^T$ are distributed normally as $X^i_{n_i} \sim N(\mu^i, \sum^i)$, $i = 0, 1$, [18], where for $k = 1, 2, \ldots, n_i$,

$$\begin{aligned}&\mu^i = [\mu^i, \mu^i, \ldots, \mu^i]^T_{1\times n_i}, \rho_s^{ii}(k) = 0, \rho_s^{01} = \mathbf{0}_{1\times n_1}, \\ &\rho_s^{10} = \mathbf{0}_{1\times n_0}, \ \sum{}^i(k,l) = \begin{cases} \sigma_i^2(1+\theta_i^2), & k=l \\ \sigma_i^2\theta_i, & |k-l|=1 \\ 0 & otherwise \end{cases},\end{aligned} \quad (83)$$

where $\mu_i = c_i$ and $\Sigma^i(k, l)$ denotes the entry in the k^{th} row and l^{th} column of the matrix Σ^i. The result follows by replacing (83) in Theorem 1.

Proof of Corollary 13

From Theorem 11, since $\alpha^0_{t_s} = \alpha^1_{t_s}$ and $\max\{n_0, n_1\} + 1 < s$, we have $2E[\varepsilon^{MA(1)_\theta}_{t_s,2n}] = G(h_\theta, l_\theta; 0)$, for any s, with $l_\theta = -k_\theta$, h_θ and k_θ defined in Corollary 12, and $G(h_\theta, -k_\theta, 0)$ defined as in (24). Similar to proof of Corollary 10, the present corollary follows by setting $n_0 = n_1 = n$ and using

$$\begin{aligned}\frac{dl_\theta}{d\theta} &= b^{-\frac{3}{2}}\frac{(c_0-c_1)\sigma^2}{n}(2\theta+2-\tfrac{2}{n}), \\ \frac{dh_\theta}{d\theta} &= -a^{-\frac{3}{2}}\frac{(c_0-c_1)\sigma^2}{4n}(2n\theta+\theta+1-\tfrac{1}{n}),\end{aligned} \quad (84)$$

where a and b are obtained from (53).

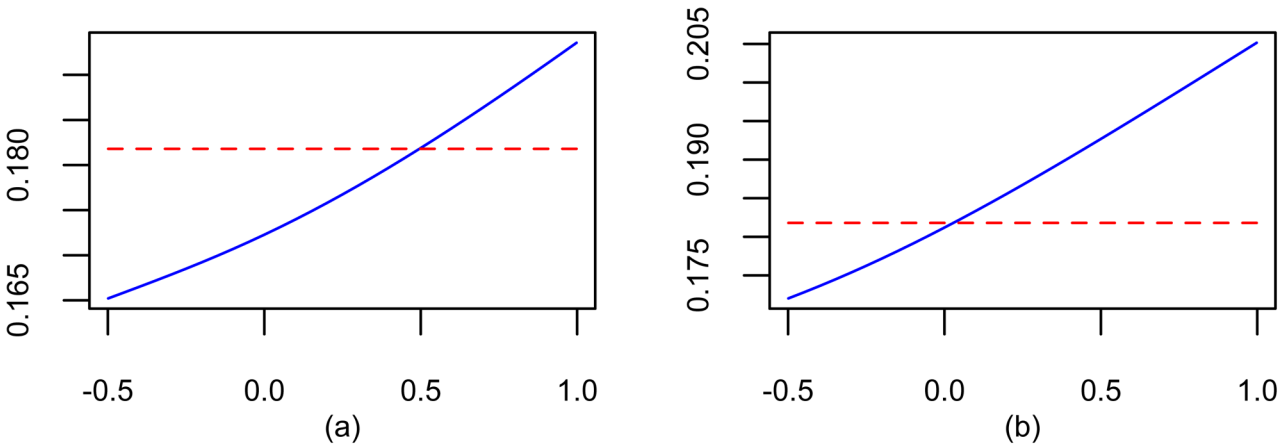

Figure 1.

(a) Expected true error of constructed classifiers in scenario *a* as a function of ρ, (b) Expected true error of constructed classifiers in scenario *b* as a function of ρ. The horizontal line shows the performance of the constructed classifier as if the samples were independent. Solid lines: dependent samples; Dashed lines: independent samples.

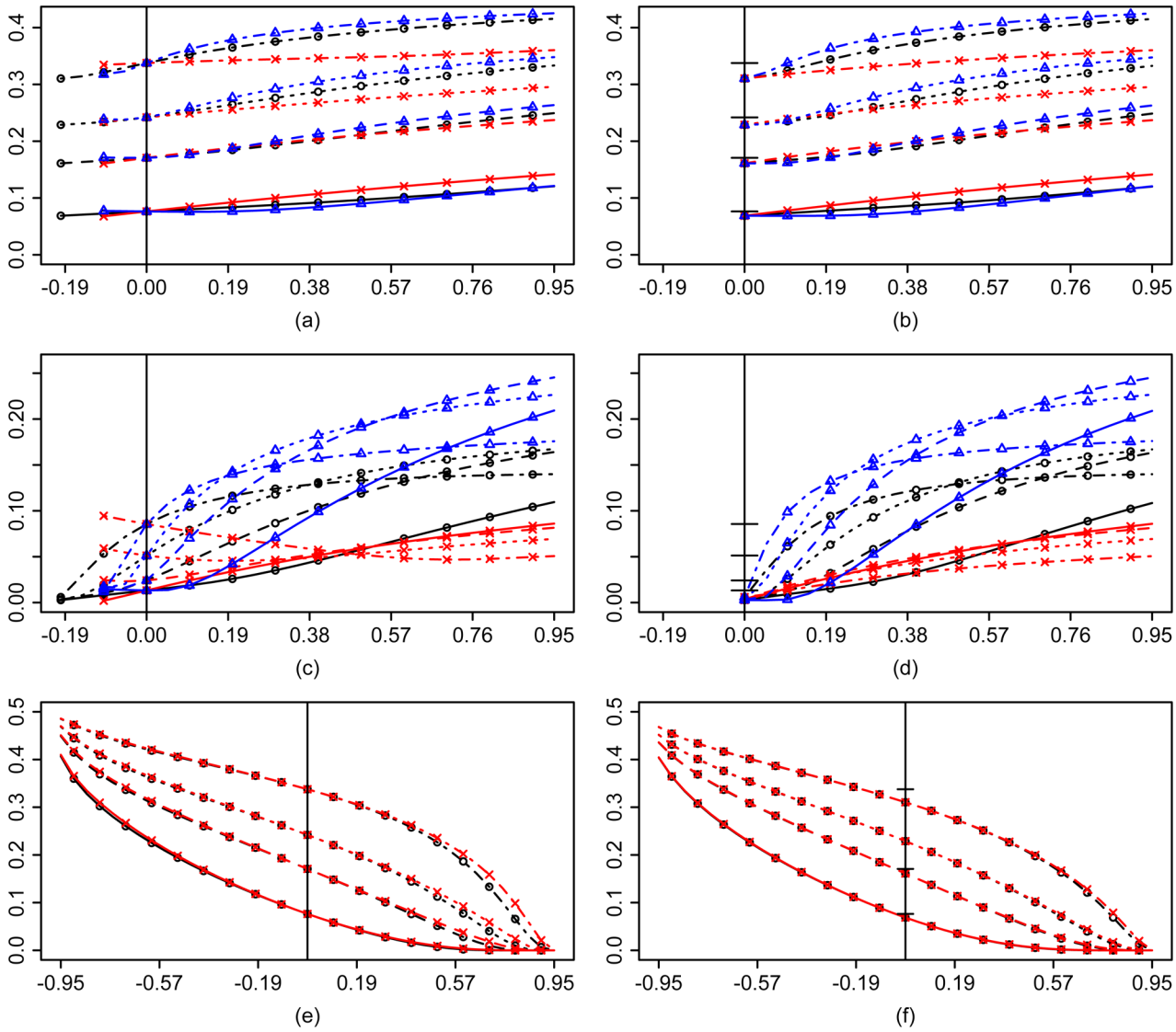

Figure 2.

Figures (a)–(d) show the exact expectation and standard deviation of LDA true error in Experiment 1 as a function of ρ_{with}. (a) Expectation for $n_0 = n_1 = 5$; (b) Expectation for $n_0 = n_1 = 25$; (c) Standard deviation for $n_0 = n_1 = 5$; (d) Standard deviation for $n_0 = n_1 = 25$; (a)–(d) plot keys: ○ := $\rho_{bet} = 0$; ×:= $\rho_{bet} = \rho_{with}$; △ = $\rho_{bet} = -\rho_{with}$; solid := $\mu_0 = 1.5$; dash := $\mu_0 = 1$; dot := $\mu_0 = 0.75$; dash-dot := $\mu_0 = 0.5$. The cross section of each curve with the vertical solid line in (a)–(d) plots shows the magnitude of the expectation/variance for i.i.d sampling situation for the corresponding scenario. The small horizontal solid lines in Figures (b) and (d) show the magnitude of expectation/variance of i.i.d situation in Figures (a) and (c), respectively. Figures (e)–(f) show the exact expectation of LDA true error of the first-order autoregressive model in Experiment 2 as a function of ψ:= $\psi_0 = \psi_1$. (a) Case of $n_0 = n_1 = 5$; (b) Case of $n_0 = n_1 = 25$; (e)–(f) plot keys: ○ := $s - n_0 = 2$; × : $s - n_0 = 10$; 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 (e)–(f) plots shows the magnitude of the expectation for i.i.d sampling situation for the corresponding scenario. The small horizontal solid lines in Figure (f) show the magnitude of expectation of i.i.d situation in Figure (e).

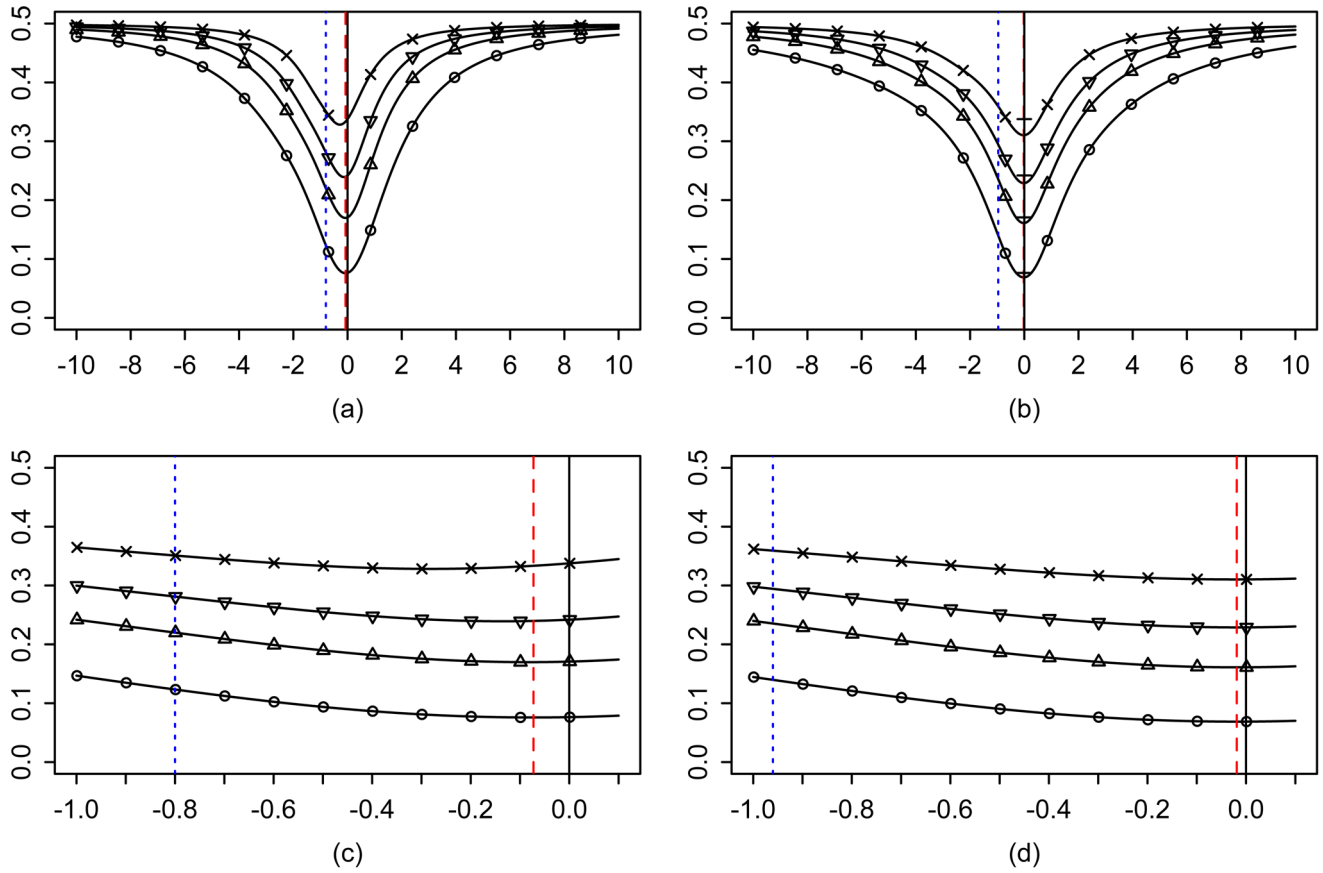

Figure 3.

Exact expectation of LDA true error of the first-order moving average model in Experiment 3 as a function of $\theta := \theta_0 = \theta_1$. (a) Expectation for $n = 5$; (b) Expectation for $n = 25$; (c) Magnification of region [−1, 0.1] in Figure (a); (d) Magnification of region [−1, 0.1] in Figure (b). Plot keys: ○:= $c_0 = 1.5$; △ : $c_0 = 1$; ▽ := $c_0 = 0.75$; × := $c_0 = 0.5$; 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. The small horizontal solid lines in Figure (b) are drawn to facilitate comparing this figure with Figure (a) at these cross sections. The left side of blue dotted line is $(-\infty, \frac{1}{n} - 1]$ region, which in (54) we proved that the expectation of true error is a decreasing function of θ. The right hand side of the red dashed line is $[\frac{\frac{1}{n}-1}{2n+1}, \infty)$ region, which the expectation is an increasing function as seen from (54).