BIOINFORMATICS

Vol. 00 no. 00 2016
Pages 1–7

An Efficient Method to Estimate the Optimum Regularization Parameter in RLDA

Daniyar Bakir [1], Alex Pappachen James [1], and Amin Zollanvari [1]

[1]Department of Electrical and Electronics Engineering, Nazarbayev University, Astana, Kazakhstan

ABSTRACT

Motivation: The biomarker discovery process in high-throughput genomic profiles has presented the statistical learning community with a challenging problem, namely learning when the number of variables is comparable or exceeding the sample size. In these settings, many classical techniques including linear discriminant analysis (LDA) falter. Poor performance of LDA is attributed to the ill-conditioned nature of sample covariance matrix when the dimension and sample size are comparable. To alleviate this problem regularized LDA (RLDA) has been classically proposed in which the sample covariance matrix is replaced by its ridge estimate. However, the performance of RLDA depends heavily on the regularization parameter used in the ridge estimate of sample covariance matrix.

Results: We propose a range-search technique for efficient estimation of the optimum regularization parameter. Using an extensive set of simulations based on synthetic and gene expression microarray data, we demonstrate the robustness of the proposed technique to Gaussianity, an assumption used in developing the core estimator. We compare the performance of the technique in terms of accuracy and efficiency to classical techniques for estimating the regularization parameter. In terms of accuracy, the results indicate that the proposed method vastly improves on similar techniques that use classical plug-in estimator. In that respect, it is better or comparable to cross-validation based search strategies while, depending on the sample size and dimensionality, being tens to hundreds times faster to compute.

Contact: amin.zollanvari@nu.edu.kz

1 INTRODUCTION

Ridge estimation is a type of shrinkage and traces back to the pioneering work of Hoerl and Kennard (Hoerl, 1962; Hoerl and Kennard, 1970a,b) on estimating regression parameters. They considered the standard linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,, \tag{1}$$

where $\mathbf{y}$ is the n dimensional observation vector, $\mathbf{X}$ is a known $n \times p$ matrix, $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_p]^T$ is a p dimensional parameter vector to be estimated, and $\boldsymbol{\varepsilon}$ is the n dimensional error vector with mean $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}_p$. If we assume $\mathbf{X}$ is a full (column) rank matrix ($p < n$), the ordinary least-square solution to this familiar linear model is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{2}$$

However, when $p > n$, the solution (2) does not exist because $\mathbf{X}^T\mathbf{X}$ becomes degenerate. Even the solution obtained by generalized inverse form of matrix $\mathbf{X}^T\mathbf{X}$ is not working well. Hoerl and Kennard (Hoerl, 1962; Hoerl and Kennard, 1970a,b) then formulated a problem in which the residual sum of squares is replaced by its ℓ_2 penalized form given by

$$L_2(\boldsymbol{\beta}) \triangleq ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + k||\boldsymbol{\beta}||^2, \tag{3}$$

where $k > 0$ denotes a penalty factor controlling the length of $\boldsymbol{\beta}$. Minimizing $L_2(\boldsymbol{\beta})$ results in the so-called ridge regression given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y}. \tag{4}$$

In this way the inverse of possibly ill-conditioned $\mathbf{X}^T\mathbf{X}$ is stabilized by adding the scalar matrix $k\mathbf{I}_p$. This idea was then used by Di Pillo (Pillo, 1976) to replace the estimate of the sample covariance matrix used in linear discriminant analysis (LDA) by its ridge estimate resulting in the so-called regularized LDA (RLDA). The goal is to improve the performance of LDA in situations where dimensionality of observations, p, is larger or comparable to the number of measurements, n. In (Pillo, 1979), Di Pillo attempts to determine the optimum value of the optimum regularization parameter in RLDA. On this Di Pillo's study, Peck and Van Ness comment that (Peck and Ness, 1982), "He found the analytical solution to this problem intractable, and so used a simulation study to choose an optimum value for k [the regularization parameter]. He concluded that if an algorithm can be found which leads to a value of k near the optimum value, then considerable improvement in the PCC [probability of correct classification] should occur".

In (Friedman, 1989), Friedman suggested the use of cross-validation in finding the optimum value of regularization parameter. In this procedure, cross-validation is used to estimate the true error of RLDA for each value of the regularization parameter selected from a pre-specified set of size 25 to 50. The estimate of the optimum regularization parameter is then the one that results in minimum cross-validation estimate of true error. Despite the computational complexity of cross-validation in such a search algorithm (e.g. see comments in (Friedman, 1989; Sharma *et al.*, 2014; Tasjudin and Landgrebe, 1998)), this approach has remained the most popular method in estimating the optimum value of regularization parameter in RLDA—for instance, see (Guo *et al.*, 2007; Bandos *et al.*, 2009; Ye *et al.*, 2006; Huang *et al.*, 2009; Ye and Xiong, 2006) to cite just a few articles.

Recently, we constructed a generalized consistent estimator of true error of RLDA. In this regard, we proposed an estimator that converges to true error in a double asymptotic sense. In this setting, the estimator converges to the actual parameter in an asymptotic scenario in which dimension and sample size

increase in a proportional manner ($n \to \infty$, $p \to \infty$, and $p/n \to J > 0$) (Zollanvari and Dougherty, 2015). In developing this estimator, we assumed that the true distributions governing the data follow multivariate Gaussian model. However, the underlying mechanism to develop the estimator was based on double asymptotics and random matrix theory, both of which suggest applicability of the estimator in non-Gaussian settings as well (see p. xii in (Girko, 1995), p. 335 in (Bai and Silverstein, 2010), and (Zollanvari, 2015)). In this work, we employ this estimator of true error in a one-dimensional search to estimate the optimum regularization parameter of RLDA. As such, we employ data taken from seven gene expression microarray studies as well as synthetically generated Gaussian and non-Gaussian data. We compare the performance (in terms of accuracy and efficiency) of the search technique that uses this estimator with similar search schemes that use cross-validation or plug-in estimators. Using an extensive set of simulations, we observe that the proposed technique is an efficient method that can outperform cross-validation based schemes in estimating the optimum regularization parameter of RLDA. Throughout this work, we use boldface lower case letters to denote a column vector. A boldface upper case letter denotes a matrix and tr[.] is the trace operator. The identity matrix of p dimension is denoted by $\mathbf{I}_p$.

2 SYSTEMS AND METHODS

2.1 RLDA Classifier

Assume a separate sampling scheme is employed: $n = n_0 + n_1$ sample points are collected to constitute the sample S in R^p, where, n, n_0 and n_1 are non-random and pre-determined and where $S_0 = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_0}\}$ and $S_1 = \{\mathbf{x}_{n_0+1}, \mathbf{x}_{n_0+2}, \ldots, \mathbf{x}_n\}$ are randomly selected from populations Π_0 and Π_1, respectively. In this two-class problem, a classifier is a function $\psi_n : R^p \to \{0, 1\}$. If ψ_n is given by $\psi_n(\mathbf{x}) = 0$ if $\mathbf{x} \in R_0$ and $\psi_n(\mathbf{x}) = 1$ if $\mathbf{x} \in R_1$, where R_0 and R_1 are measurable sets partitioning the sample space, then the true error of ψ_n, denoted by ε, is defined to be the probability of misclassification,

$$\varepsilon = \alpha_0 \int_{R_1} f(\mathbf{x}|0)\, d\mathbf{x} + \alpha_1 \int_{R_0} f(\mathbf{x}|1)\, d\mathbf{x} = \alpha_0\varepsilon_0 + \alpha_1\varepsilon_1, \tag{5}$$

where α_i is the prior probability for class i, ε_i is the error contributed by class i, and $f(\mathbf{x}|0)$ and $f(\mathbf{x}|1)$ are the class-conditional densities governing Π_0 and Π_1, respectively. Separate sampling is very common in biomedical applications, where data from two classes are collected without reference to the other class, for instance, when discriminating two types of tumors or when distinguishing a normal from a pathological phenotype. With separate sampling, the prior probabilities α_i cannot be estimated from the sample, an issue with a long history in the study of LDA (Anderson, 1951). Both classification rules (Esfahani and Dougherty, 2014) and error estimation rules (Braga-Neto *et al.*, 2014) need to be adjusted for separate sampling rather than use their usual random-sampling definitions; otherwise, they suffer performance degradation. The adjustment requires that α_0 and α_1 be known, as assumption made in this study. In our case the adjustment is straightforward because it simply means that we directly use α_0 and α_1 rather than their random-sampling estimates $\frac{n_0}{n}$ and $\frac{n_1}{n}$. In practice, the salient point is that, given n, n_0 and n_1 are chosen so that $\frac{n_0}{n}$ is as close to α_0 as possible (Esfahani and Dougherty, 2014).

Assuming Π_i follows a multivariate Gaussian distribution $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, for $i = 0, 1$, where $\boldsymbol{\Sigma}$ is the common nonsingular covariance matrix of both class, replacing the unknown mean and the covariance matrix of classes in Bayes rule (optimum classifier) results in LDA, which is characterized by Anderson's statistics,

$$W^{LDA}(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{C}, \mathbf{x}) = \left(\mathbf{x} - \frac{\bar{\mathbf{x}}_0 + \bar{\mathbf{x}}_1}{2}\right)^T \mathbf{C}^{-1}(\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1), \tag{6}$$

where $\bar{\mathbf{x}}_0 = \frac{1}{n_0}\sum_{\mathbf{x}_l \in S_0} \mathbf{x}_l$ and $\bar{\mathbf{x}}_1 = \frac{1}{n_1}\sum_{\mathbf{x}_l \in S_1} \mathbf{x}_l$ are the sample means for classes 0 and 1 respectively, and $\mathbf{C}$ is the pooled sample covariance matrix,

$$\mathbf{C} = \frac{(n_0 - 1)\,\mathbf{C}_0 + (n_1 - 1)\,\mathbf{C}_1}{n_0 + n_1 - 2}, \tag{7}$$

where

$$\mathbf{C}_i = \frac{1}{n_i - 1}\sum_{\mathbf{x}_l \in S_i} (\mathbf{x}_l - \bar{\mathbf{x}}_i)(\mathbf{x}_l - \bar{\mathbf{x}}_i)^T. \tag{8}$$

In this work we consider a form of RLDA classifier that is obtained by using ridge estimators of the inverse covariance matrix in W^{LDA}; that is, by using $(I + \gamma\mathbf{C})^{-1}$ and $\gamma > 0$, in (6), which yields

$$W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{C}, \mathbf{x}) = \left(\mathbf{x} - \frac{\bar{\mathbf{x}}_0 + \bar{\mathbf{x}}_1}{2}\right)^T \mathbf{H}(\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1), \tag{9}$$

where

$$\mathbf{H} = (\mathbf{I}_p + \gamma\mathbf{C})^{-1}. \tag{10}$$

The designed RLDA classifier is then given by

$$\psi_n^{RLDA}(\mathbf{x}) = \begin{cases} 1, & \text{if } W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{C}, \mathbf{x}) \le c \\ 0, & \text{if } W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{C}, \mathbf{x}) > c \end{cases}, \tag{11}$$

where $c = \log\frac{1-\alpha_0}{\alpha_0}$.

2.2 RLDA True Error, Optimum Regularization, and Their Estimates

The true error of ψ_n^{RLDA} is given by (5). Given sample S_n, for $i = 0, 1$,

$$\varepsilon_i = P((-1)^i W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{C}, \mathbf{x}) \le (-1)^i c \mid \mathbf{x} \in \Pi_i, \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{C}). \tag{12}$$

Under the multivariate Gaussian model, we have

$$\varepsilon_i = \Phi\left(\frac{(-1)^{i+1} G\big(\boldsymbol{\mu}_i, \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{H}\big) + (-1)^i c}{\sqrt{D\big(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{H}, \boldsymbol{\Sigma}\big)}}\right), \tag{13}$$

where $\Phi(.)$ denotes the cumulative distribution function of a standard normal random variable and

$$\begin{aligned} G\big(\boldsymbol{\mu}_i, \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{H}\big) &= \left(\boldsymbol{\mu}_i - \frac{\bar{\mathbf{x}}_0 + \bar{\mathbf{x}}_1}{2}\right)^T \mathbf{H}(\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1), \\ D\big(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{H}, \boldsymbol{\Sigma}\big) &= (\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)^T \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}(\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1). \end{aligned} \tag{14}$$

Given training data, the optimal choice of γ is the value of γ, which minimizes the overall true error ε as defined by (5) and

(13); to wit, $\gamma^{opt} = \underset{\gamma}{\operatorname{argmin}}\, \varepsilon$. However, true error depends on unknown population parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}$, which must be estimated from training data. As such, the optimum regularization parameter depends on unknown distributional parameters and must be estimated from data as well. Even with the assumption of knowing the true distributional parameters, γ^{opt} is the solution of a non-linear equation that needs to be solved numerically. To see the latter statement and for simplicity of presentation, let $\alpha_i = 1$ and $\alpha_{1-i} = 0$, $i = 0, 1$, which means $\gamma^{opt} = \underset{\gamma}{\operatorname{argmin}}\, \varepsilon = \underset{\gamma}{\operatorname{argmin}}\, \varepsilon_i$. By taking the derivative of ε_i defined in (13) with respect to γ, setting the derivative to zero, and after some tedious but straightforward algebraic manipulations we observe that γ^{opt} is the unique positive solution of the following equation,

$$\frac{G\big(\boldsymbol{\mu}_i, \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{HCH}\big)}{D\big(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{H}, \mathbf{HC\Sigma}\big)} = \frac{G\big(\boldsymbol{\mu}_i, \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{H}\big)}{D\big(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{H}, \boldsymbol{\Sigma}\big)}, \tag{15}$$

where dependency of equation on γ is via $\mathbf{H}$ defined in (10). The non-linearity of the equation makes a closed form expression of γ^{opt} hopeless. As such, a range search strategy is a feasible path forward.

The objective in the range search is to determine the γ that minimizes the estimate of true error of RLDA. In this regard, a classical estimate of true error is obtained by replacing the unknown parameters by their sample estimate, resulting in standard plug-in estimator of true error, which is given by (McLachlan, 2004)

$$\hat{\varepsilon}_i^P = \Phi\left(\frac{(-1)^{i+1} G\big(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{H}\big) + (-1)^i c}{\sqrt{D\big(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{H}, \mathbf{C}\big)}}\right). \tag{16}$$

It is straightforward to see that for fixed p, as $n_i \to \infty$, we have $\bar{\mathbf{x}}_i \to \boldsymbol{\mu}_i$ and $\mathbf{C} \to \boldsymbol{\Sigma}$, and therefore, $\hat{\varepsilon}_i^P \xrightarrow{\text{P}} \varepsilon_i^{RLDA}$ where $\xrightarrow{\text{P}}$ denotes convergence in probability.

In (Zollanvari and Dougherty, 2015), we proposed the following estimator for true error of RLDA:

$$\hat{\varepsilon}_i^D = \Phi\left(\frac{(-1)^{i+1} G\big(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{H}\big) + \frac{(n_0+n_1-2)\hat{\delta}}{n_i} + (-1)^i c}{\sqrt{(1+\gamma\hat{\delta})^2 D\big(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{H}, \mathbf{C}\big)}}\right), \tag{17}$$

where

$$\hat{\delta} = \frac{\frac{p}{n_0+n_1-2} - \frac{\text{tr}[\mathbf{H}]}{n_0+n_1-2}}{\gamma\left(1 - \frac{p}{n_0+n_1-2} + \frac{\text{tr}[\mathbf{H}]}{n_0+n_1-2}\right)}. \tag{18}$$

Using random matrix theory and under double asymptotic conditions, the estimator (17) converges (almost surely) to true error. The double asymptotic conditions are mainly characterized by $n_0 \to \infty, n_1 \to \infty, p \to \infty$, with the assumption that the following limits exist: $\frac{p}{n_0} \to J_0 > 0, \frac{p}{n_1} \to J_1 > 0$, and $\frac{p}{n_0+n_1} \to J < \infty$. Nevertheless, the readers are referred to (Zollanvari and Dougherty, 2015) for the complete list of conditions used in developing (17).

We use the following protocol to estimate γ^{opt} using a set of banchmark gene expression datasets and, at the same time, compare the performance of the proposed search strategy based on various estimators of error. The estimators that we use are 5-fold cross-validation with 5 repetitions (CV5F-5R), leave-one-out (loo), plug-in ($\hat{\varepsilon}^P$) available from (16), and our proposed double-asymptotic estimator $\hat{\varepsilon}_i^D$ available from (17). The experiments on real data and synthetic data are essentially similar except that in real-data experiments we employ t-test feature selection to reduce the dimensionality to $p = 50$ and $p = 150$.

Protocol (Real Data):

- **Step I**: Let r denote the ratio of the total number of sample points in class 0 to the total amount in class 1 in the full dataset. Let n_{Full} denote the sample size in the full dataset. Fix a value $n < n_{Full}$ and let it be the number of training sample points that are randomly taken out of the whole dataset such that $n = n_0 + n_1$ with n_i being the number of training sample points in class i. We choose $n_0 = \lfloor r n_1 \rfloor$, where $\lfloor . \rfloor$ is the floor function. This practice resembles a random sampling scheme in which $\alpha_0 \approx \frac{n_0}{n}$ and $\alpha_1 \approx \frac{n_1}{n}$. Therefore, we use these values of α_i to find the overall error rate from (5) and the held-out samples. In order to set aside enough sample points for testing (i.e. the $n_{Full} - n$ held-out sample), we restrict the training sample size to $n \in [30, 100]$.
- **Step II**: For a prescribed value of regularization parameter γ in a prescribed range, design the RLDA classifier by (9). We discretize the range with the exponential function $\left(1000^{\frac{1}{10}}\right)^i$ for $i = \{-10, -9, -8, \ldots, 10\}$ that covers values from 0.001 to 1000. The above exponential function has been chosen to improve the efficiency of the search. This choice seems to be a reasonable one because a small perturbation in large values of γ is a smaller relative change with respect to a similar perturbation in small values of γ. This implies that the effect of the former perturbation in changing the true error of the classifier may not be as large as the latter perturbation (although in terms of magnitude both perturbation are the same). In other words, for large values of γ having a fine discretization is not as critical as small values.
- **Step III**: For each value of γ in the prescribed set of points, estimate the error of the designed classifier using as estimator of error (CV5F-5R, loo, $\hat{\varepsilon}^P$, and $\hat{\varepsilon}^D$). Obtain the holdout estimate of the true error (taken as the true error) from the test data.
- **Step IV**: The estimate of the optimum γ is the γ which results in the smallest error estimate on the prescribed range of γ. For the estimated optimum γ record the value of true error (available from step III).
- **Step V**: Repeat Steps I-IV, 500 times for each n and determine the average expected error of RLDA.

3 RESULTS AND DISCUSSION

Based on the protocols described in Section 2, we have performed a set of experiments employing both synthetic models and gene expression microarray data to examine the performance of the search scheme based on various estimators. First, we consider seven publicly available datasets on: breast cancer (van de Vijver *et al.*, 2002), pediatric acute lymphoblastic leukemia (Yeoh *et al.*, 2002), hepatocellular carcinoma (Chen *et al.*, 2004), toxicants response on rats (Natsoulis *et al.*, 2005), diffuse large B-cell lymphoma

(Rosenwald *et al.*, 2002), node-negative breast cancer (Desmedt *et al.*, 2007), and acute myeloid leukemia (Valk *et al.*, 2004). Table 1 provides a summary of these datasets, including the total number of genes and sample size. For a description of the data preparation, the readers are referred to the Supplementary Materials. Fig. 1 (Fig. S1) show the expected true and estimate of error for RLDA classifier as a function of regularization parameter γ for different number of sample points ranging from 30 to 100 chosen from datasets listed in Table 1 with $p = 50$ ($p = 150$). This leaves us with 8 (sample sizes) $\times$ 7 (datasets) $\times$ 2 (dimensionalities)=112 experiments on real data. As seen in the far right column of these figures, for each sample size, the true error of classifier decreases as a function of γ and then increases for increasing γ with the optimal γ corresponding to the minimum true error at the bottom of the valley. In this regard, in all experiments such a "peaking phenomenon" occurs in the pre-specified range of $\gamma \in [0.001, 1000]$ with 75% of times (84 out of 112) happening in the range $[0.1, 100]$. Notice that this peaking phenomenon is also observed in curves of estimated errors (columns 1-3 in Fig. 1 and Fig. S1) except for the plug-in estimator, suggesting that plug-in is not a good estimator of the optimum γ.

Table 1. Microarray studies used in this work

Dataset	Features	n_0/n_1
(Chen *et al.*, 2004)	$10,237$	$75/82$
(Desmedt *et al.*, 2007)	22215	$98/77$
(Natsoulis *et al.*, 2005)	$8,491$	$120/61$
(Rosenwald *et al.*, 2002)	$5,013$	$114/89$
(Valk *et al.*, 2004)	22215	$116/157$
(van de Vijver *et al.*, 2002)	$10,237$	$180/115$
(Yeoh *et al.*, 2002)	$5,077$	$149/99$

Fig. 2-(a) to (n) show the expected true error of RLDA classifier designed using the estimate of the optimum γ (the γ that results in the minimum *estimated error* in Fig. 1 and Fig. S1) obtained from various estimators as a function of sample size on each dataset. We observe that an RLDA classifier designed by double asymptotic estimator $\hat{\varepsilon}^D$ has a better or comparable performance to RLDA classifiers constructed using plug-in, CV5F-5R, and loo estimators. At the same time, we have to note that to compute $\hat{\varepsilon}^D$, we only need to evaluate the closed-form expression presented in (17). Consequently, $\hat{\varepsilon}^D$ is tens to hundreds of times faster to compute than cross-validation estimators. To illustrate this point, we have plotted the ratio of average time it takes to compute CV5F-5R and leave-one-out estimators to the time it takes to compute $\hat{\varepsilon}^P$ and $\hat{\varepsilon}^D$ estimators in experiments related to (Chen *et al.*, 2004) (see Fig. 3). The actual average compute time is presented in the Supplementary Materials Section 6.

Note that the pre-specified range of γ is important to obtain a realistic view of the performance of estimators. For example, if we limit the search range of γ to $[0.1, 100]$, then in the Natsoulis' experiment, the classical plug-in estimator $\hat{\varepsilon}^P$, which is not expected to have a good performance in small-sample situations, outperforms all other estimators (see Fig. S2). This behavior is because in this dataset for all examined sample sizes the optimum regularization parameter is larger than or close to the upper limit of the range of $\gamma \in [0.1, 100]$. This can be seen from the figure on the third row, fifth column in Fig. 1. At the same time in all datasets, $\hat{\varepsilon}^P$ points to the upper bound of the range as the estimate of the optimum regularization parameter, which in the Natsoulis' experiment happens to be closer to the actual optimum regularization parameter (see the plot in the third row, fourth column of Fig. 1).

We also used synthetic data to compare the performance of estimators in estimating optimum γ. Figure 2-(o) to (t) show the results for a wide range of Bayes (optimum) error and $p = 20$ for data taken from Gaussian and skew-normal distributions. For the complete set of results along with the protocol used for synthetic experiments see Section 4 and 5 in the Supplementary Materials. In almost all synthetic experiments, $\hat{\varepsilon}^D$ uniformly outperforms other estimators of γ.

The efficiency of the proposed procedure is a direct consequence of having a closed form for the core estimator that we use in the search. The good performance is due to convergence of the core estimator to true error in a double asymptotic regime. Classically, the notion of statistical consistency guarantees the performance of an estimator in situations where the number of measurements unboundedly increases for a fixed dimensionality ($n \to \infty$, p fixed). In a finite sample operating regime, this implies that in order to expect an acceptable performance from an estimator, we need to have many more sample points than variables. However, in a double asymptotic regime the magnitude of p and n are kept comparable ($p/n \to J > 0$ with J being an arbitrary number) and, as a result, we generally expect an acceptable performance of developed estimators in a wide range of dimension and sample size. We note that both cross-validation and plug-in estimators are statistically consistent in a classical sense while the core estimator that we use in the search is a consistent estimator in a double asymptotic sense.

4 CONCLUDING REMARKS

A recently proposed estimator of true error of RLDA based on double asymptotics is used in a one-dimensional search to optimize the performance of the classifier in terms of regularization parameter. While in developing the core estimator used in the search we have assumed the Gaussianity of the data, the underlying mechanism to develop the estimator is based on random matrix theory. The universality principle of random matrix theory though suggests applicability of developed estimators in non-Gaussian settings as well (see p. xii in (Girko, 1995), p. 335 in (Bai and Silverstein, 2010), and (Zollanvari, 2015)). In this work we conducted an extensive set of simulations using both synthetic and gene expression microarray data to compare the performance of our technique in terms of expected error of the constructed RLDA and the compute time to similar search schemes that use classical error estimators (5-fold cross-validation with 5 repetitions, leave-one-out, and plug-in estimator). We observe that the proposed technique is tens to hundreds of times faster than cross-validation to compute, while at the same time results in a comparable or better classification accuracy of the constructed RLDA. The good accuracy of the proposed technique on non-Gaussian real data and synthetic data used in this study confirms robustness of the estimator to non-Gaussianity of data. The next natural step in this line of work is to

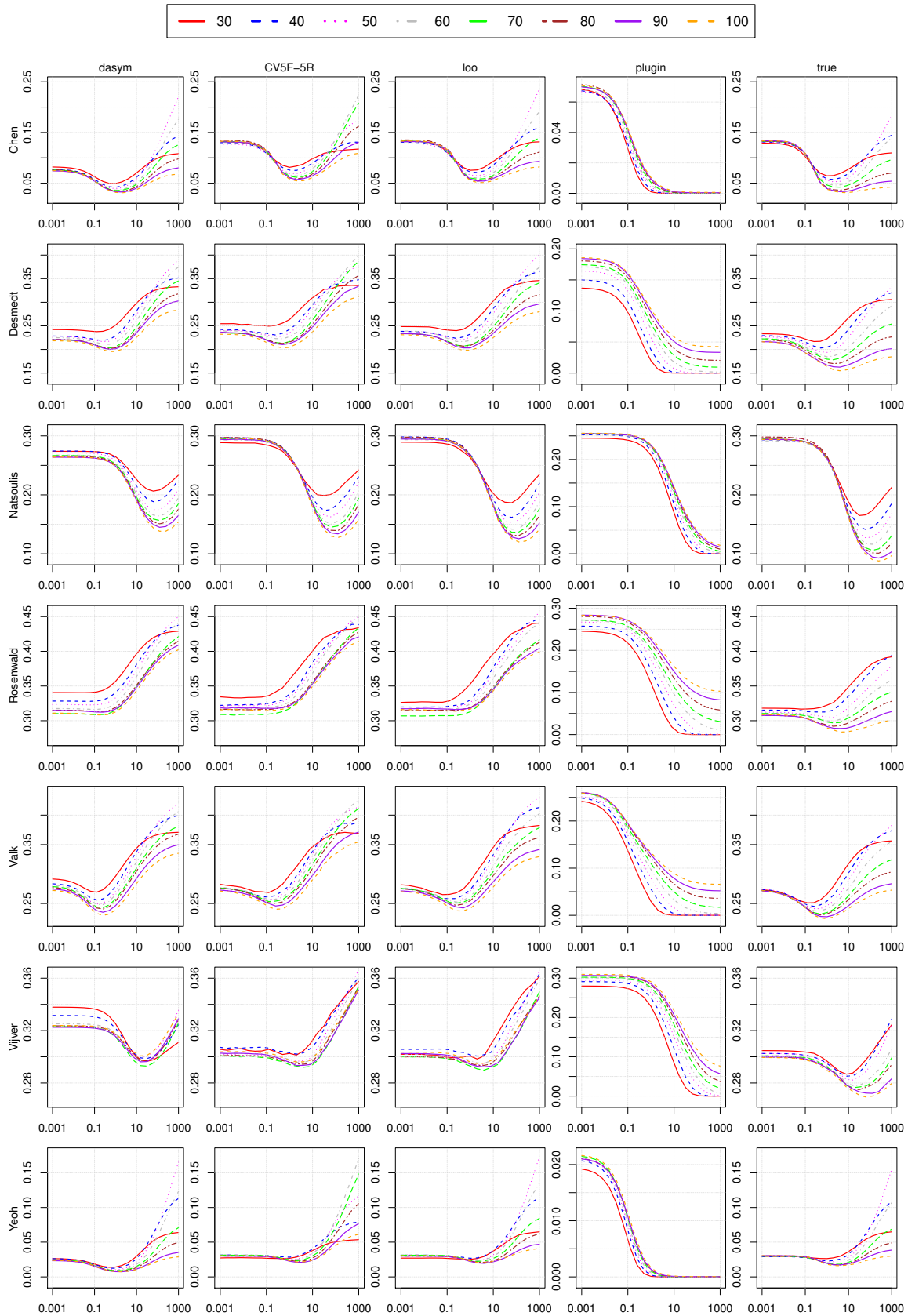

Fig. 1: Expected estimated and true error (vertical axis) as a function of regularization parameter in logarithmic scale (horizontal axis) for real datasets and feature size $p = 50$. Columns from left to right: the double asymptotic estimator $\hat{\varepsilon}^D$ (identified by dasym-est), CV5F-5R, leave-one-out (identified by loo), the plug-in estimator $\hat{\varepsilon}^P$, and the true error. Rows from top to bottom: (Chen *et al.*, 2004), (Desmedt *et al.*, 2007), (Natsoulis *et al.*, 2005), (Rosenwald *et al.*, 2002), (Valk *et al.*, 2004), (van de Vijver *et al.*, 2002), and (Yeoh *et al.*, 2002) studies. The x-axis denotes the regularization parameter ranging from 10^{-3} to 10^{3}. Note that the range of vertical axis for plug-in estimator differs from others estimators due to substantial difference between magnitude of plug-in estimates from others.

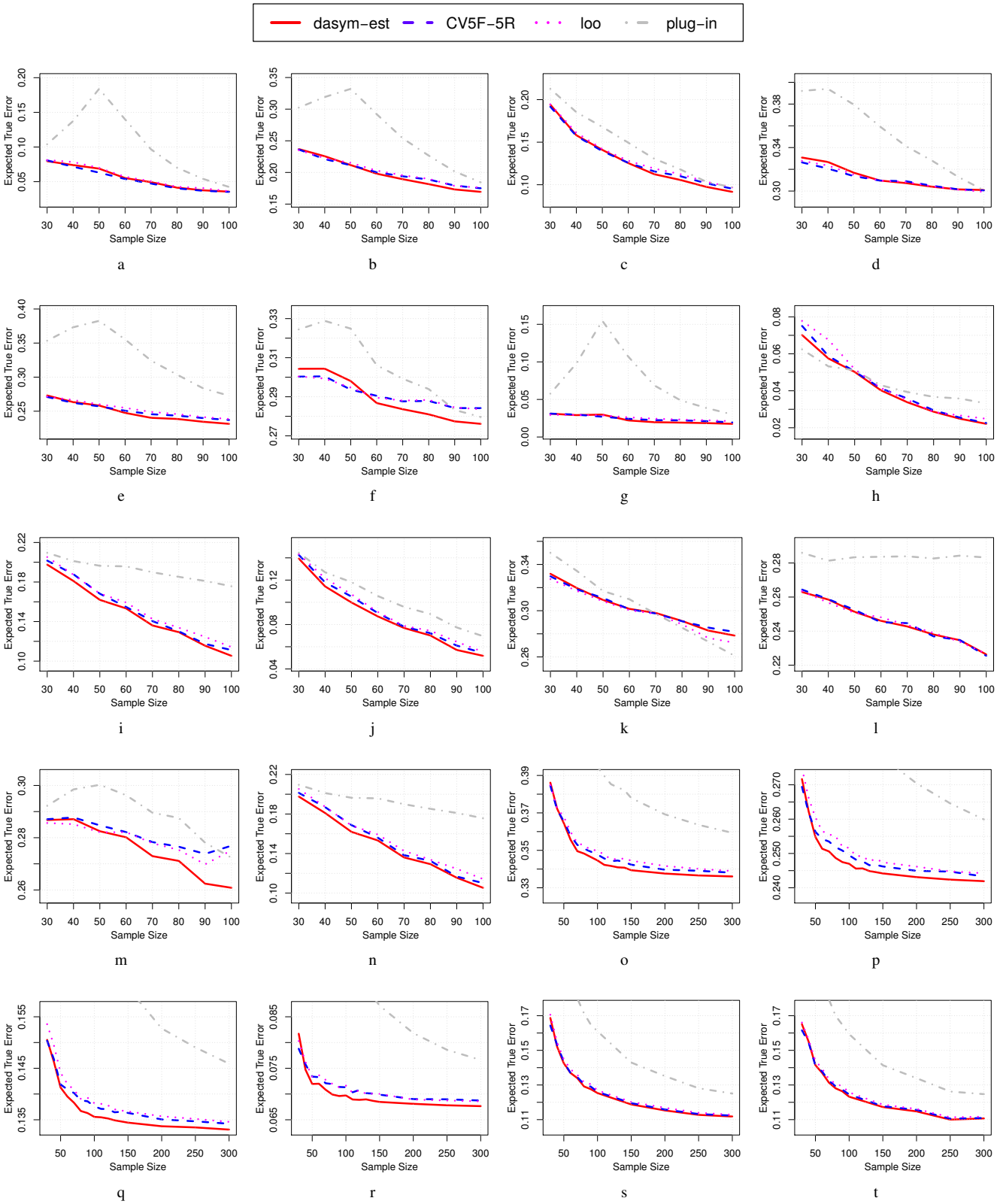

Fig. 2: The performance (expected true error) of RLDA classifiers with regularization parameter determined using different estimators of true error versus sample size for different dimensionality on real and synthetic data. The estimators used are the double asymptotic estimator $\hat{\varepsilon}^D$ (identified by dasym-est), CV5F-5R, leave-one-out (identified by loo), and the plug-in estimator $\hat{\varepsilon}^P$. Plots (a) to (n) show the results of experiments on real data; (a) to (g) and (h) to (n) show results for $p = 50$ and $p = 150$, respectively. (a) and (h): (Chen *et al.*, 2004); (b) and (i): (Desmedt *et al.*, 2007); (c) and (j): (Natsoulis *et al.*, 2005); (d) and (k) (Rosenwald *et al.*, 2002); (e) and (l) (Valk *et al.*, 2004); (f) and (m) (van de Vijver *et al.*, 2002); and (g) and (n) (Yeoh *et al.*, 2002) studies. Plots (o) to (r) show the results for synthetic data and $p = 20$: (o), (p), (q), and (r) correspond to Gaussian data and Bayes error = 0.332, 0.239, 0.131, 0.066, respectively whereas (s) and (t) correspond to skewed normal distribution with a "distance" 2 and skewness factor $\alpha = 2, 4$, respectively (see Supplementary Material Section 4 for more information on simulations and parameters regarding skew-normal distribution).

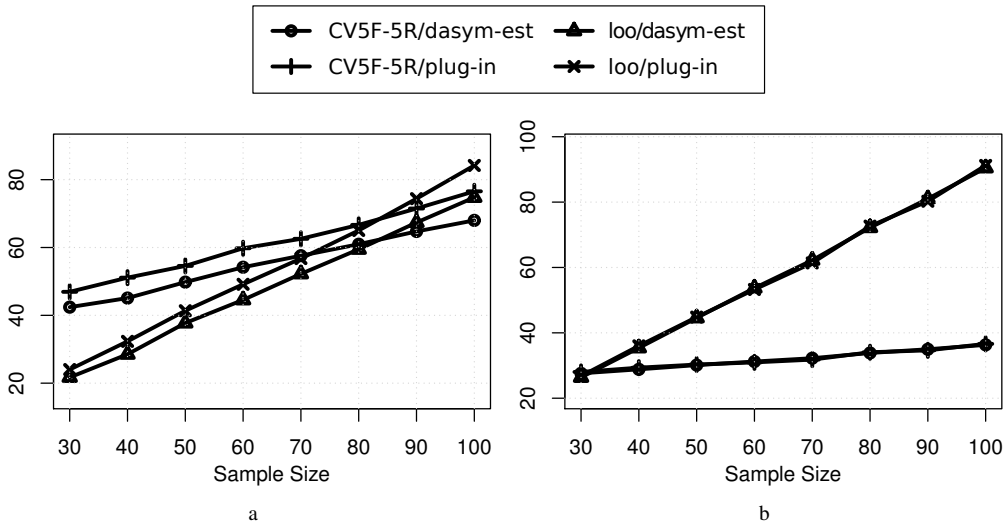

Fig. 3: The ratio of average compute time of CV5F-5R and leave-one-out estimators to average compute time of $\hat{\varepsilon}^P$ and $\hat{\varepsilon}^D$ estimators versus sample size: (a) $p = 50$; (b) $p = 150$. See Supplementary Material Section 6 for the actual compute time in terms of seconds on a personal computer.

estimate the RLDA regularization parameter that minimizes the area under the ROC curve.

REFERENCES

Anderson, T. (1951). Classification by multivariate analysis. *Psychometrika*, **16**(1), 31–50.

Bai, Z. D. and Silverstein, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*. Springer.

Bandos, T. V., Bruzzone, L., and Camps-Valls, G. (2009). Classification of hyperspectral images with regularized linear discriminant analysis. *IEEE Trans. Geosci. Remote Sens.*, **47**, 862–873.

Braga-Neto, U., Zollanvari, A., and Dougherty, E. (2014). Cross-validation under separate sampling: strong bias and how to correct it. *Bioinformatics*, **30**(23), 3349–3355.

Chen, X., Higgins, J., Cheung, S. T., Li, R., Mason, V., and *et al.* (2004). Novel endothelial cell markers in hepatocellular carcinoma. *Modern Pathol.*, **17**, 1198–1210.

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., and *et al.* (2007). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. *Clin. Cancer Res.*, **13**, 3207–3214.

Esfahani, M. S. and Dougherty, E. R. (2014). Effect of separate sampling on classification accuracy. *Bioinformatics*, **30**(2), 242–250.

Friedman, J. (1989). Regularized discriminant analysis. *J. Amer. Stat. Assoc.*, **84**, 165–175.

Girko, V. L. (1995). *Statistical Analysis of Observations of Increasing Dimension*. Kluwer Academic Publishers, Dordrecht.

Guo, Y., Hastie, T., and Tibshirani, R. (2007). Regularized discriminant analysis and its application in microarrays. *Biostat.*, **8**, 86–100.

Hoerl, A. E. (1962). Application of ridge analysis to regression problems. *Chemical Engineering Progress*, **58**, 54–59.

Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, **12**, 69–82.

Hoerl, A. E. and Kennard, R. W. (1970b). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–59.

Huang, D., Quan, Y., He, M., and Zhou, B. (2009). Comparison of linear discriminant analysis methods for the classification of cancer based on gene expression data. *Journal of Experimental & Clinical Cancer Research*, **28**, 1–8.

McLachlan, G. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.

Natsoulis, G., Ghaoui, L. E., Lanckriet, G. R., Tolley, A. M., Leroy, F., and *et al.* (2005). Classification of a large microarray data set: algorithm comparison and analysis of drug signatures. *Genome Res.*, **1**, 724–736.

Peck, R. and Ness, J. V. (1982). The use of shrinkage estimators in linear discriminant analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, **4**, 409–424.

Pillo, P. J. D. (1976). The application of bias to discriminant analysis. *Communications in Statistics - Theory and Methods*, **5**, 843–854.

Pillo, P. J. D. (1979). Biased discriminant analysis: Evaluation of the optimum probability of misclassification. *Communications in Statistics - Theory and Methods*, **8**, 1447–1457.

Rosenwald, A., Wright, G., Chan, M. C., Connors, J. M., Campo, E., and *et al.* (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New Eng. J. Med.*, **346**, 1937–1947.

Sharma, A., Paliwal, K. K., Imoto, S., and Miyano, S. (2014). A feature selection method using improved regularization discriminant analysis. *Machine Vision and Applications*, **25**, 775–786.

Tasjudin, S. and Landgrebe, D. A. (1998). Covariance estimation for limited training samples. In *Proc. Geoscience and Remote Sensing Symposium*, pages 2688–2690.

Valk, P. J., Verhaak, R. G., Beijen, M. A., Erpelinck, C. A., Barjesteh, S., and *et al.* (2004). Prognostically useful gene-expression profiles in acute myeloid leukemia. *New Eng. J. Med.*, **350**, 1617–1628.

van de Vijver, M., He, Y., and et. al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999–2009.

Ye, J. and Xiong, T. (2006). Computational and theoretical analysis of null space and orthogonal linear discriminant analysis. *Journal of Machine Learning Research*, **7**, 1183–1204.

Ye, J., Janardan, R., Cherkassky, V., Xiong, T., Bi, J., and Kambhamettu, C. (2006). Efficient model selection for regularized linear discriminant analysis. In *Proc. 15th ACM International Conf. on Information and Knowledge Management*, pages 532–539.

Yeoh, E. J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., and *et al.* (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, **1**, 133–143.

Zollanvari, A. (2015). High-dimensional statistical learning: Roots, justifications, and potential machineries. *Cancer Informatics*, **5**, 109–121.

Zollanvari, A. and Dougherty, E. R. (2015). Generalized consistent error estimator of linear discriminant analysis. *IEEE Trans. Sig. Proc.*, **63**, 2804–2814.