Does Modeling Lead to More Accurate Classification?: A Study of Relative Efficiency in Linear Classification

Yoonkyung Lee and Rui Wang [*]

Abstract

Classification arises in a wide range of applications. A variety of statistical tools have been developed for learning classification rules from data. Understanding of their relative merits and comparisons help users to choose a proper method in practice. This paper focuses on theoretical comparison of model-based classification methods in statistics with algorithmic methods in machine learning in terms of the error rate. Extending Efron's comparison of logistic regression with linear discriminant analysis (LDA) under the normal setting, we contrast such algorithmic methods as the support vector machine (SVM) and boosting with the LDA and logistic regression and study their relative efficiencies in reducing the error rate based on the limiting behavior of the classification boundary of each method. We show that algorithmic methods are generally less effective than model-based methods in the normal setting. In particular, loss of efficiency in error rate is typically about 33 to 60% for the SVM and 50 to 80% for boosting when compared to the LDA. However, a smooth variant of the SVM is shown to be even more efficient than logistic regression. In addition to the theoretical study, we present results from numerical experiments under various settings for comparisons of finite-sample performance and robustness to mislabeling and model misspecification.

Key words: Boosting; Classification; Efficiency; Error Rate; LDA; Logistic Regression; Mislabeling; Robustness; SVM

1 Introduction

Classification arises in applications from diverse domains, for example, speech recognition, spam filtering, fraud detection, and medical diagnosis. A variety of statistical tools have been developed for learning a classification (or discrimination) rule with low error rates over novel cases. To name a few, Fisher's linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) are classical examples of a discriminant rule in the statistics literature, and modern statistical tools include classification trees, logistic regression, neural networks, and kernel density based methods. For reference to classification in general, see

[*] Yoonkyung Lee is Associate Professor, Department of Statistics, The Ohio State University, Columbus, OH 43210 (E-mail: yklee@stat.osu.edu). Rui Wang is Postdoctoral Fellow, Department of Statistics, The Ohio State University. (E-mail: wang.810@buckeyemail.osu.edu). Lee's research was supported in part by NSA Grant H98230-10-1-0202 and NSF Grant DMS-12-09194.

Hastie et al. (2001); Duda et al. (2000); McLachlan (2004) and Devroye et al. (1996). More recent additions to the data analyst's toolbox for classification are the support vector machine (SVM) (Vapnik 1998; Cristianini and Shawe-Taylor 2000; Schölkopf and Smola 2002), boosting (Freund and Schapire 1997), and other margin-based methods generally dubbed *large-margin classifiers*. They have drawn considerable attention in machine learning for the last decade or so, and been successfully used in many applications of data mining, engineering, and bioinformatics; for instance, hand-written digit recognition, text categorization, and cancer diagnosis with genomic biomarkers.

Traditionally, in statistics, modeling approach to classification has been prevalent, where the underlying probability model that generates data is estimated first, and then a discrimination rule is derived from the estimated model. Logistic regression, LDA, QDA and other density based methods exemplify the approach. In contrast, in machine learning, algorithmic approach is more common, where one aims at direct minimization of the error rate without estimating a probability model explicitly by employing a convex surrogate criterion of the misclassification count (0-1). The latter yields non-probability model based methods such as SVM, boosting and other large margin classifiers.

In modern data analysis where typically high dimensional attributes are involved, refined statistical modeling may not be as tractable as in the classical setting. Also, in parallel, computational efficiency has become an ever more important factor in the applicability of a method. These aspects render algorithmic methods practically viable in a wide range of applications. The contrast between the model-based methods and algorithmic methods has also brought many interesting theoretical developments. For instance, Zhang (2004); Bartlett et al. (2006); Lin (2002) and Steinwart (2005) study the Bayes risk consistency of algorithmic approach with a convex surrogate loss and provide proper conditions for the surrogate loss to ensure the risk consistency in the binary classification problem. In particular, the hinge loss for SVM and the exponential loss for boosting are shown to be properly calibrated for the Bayes risk consistency. These results suggest that at least in terms of risk consistency, there is no appreciable difference between the two approaches theoretically. They also confirm the common belief in machine learning that formal modeling may not be necessary for classification or pattern recognition as less is required for risk consistency than in regression.

As a practical question, whether the 'soft' classification approach in statistics is more appropriate than the 'hard' classification approach in machine learning depends largely on the context of applications. Certainly, in some applications, accurate estimation of the class conditional probability given the attributes is required for making better decisions than just prediction of a likely outcome.

However, as a theoretical question, comparison of the two approaches remains open to investigation. To the best of our knowledge, their relative merits and efficiency have not been rigorously examined on the theoretical basis. Given the differences in the two paradigms of modeling versus prediction, a basic question we pose here is whether probability modeling leads to more efficient use of data in reducing the error rate than the algorithmic approach, and if so, how much efficiency is gained by modeling.

To simplify the question, we examine the effect of modeling on the error rate analytically in the normal distribution setting by computing the asymptotic relative efficiency (ARE) of various classification methods ranging from the full modeling approach of the LDA to the purely algorithmic procedure of the SVM in reducing the classification error rate (see

Section 2.3 for the definition of ARE). Drawing on Efron's classical framework for comparison of the LDA with logistic regression (Efron 1975), we present similar analysis and large-sample comparison for some of popular machine learning methods. In doing so, we use the asymptotic theory of M-estimators to characterize the limiting distribution of a discriminant function and the associated error rate for methods that are defined through convex loss criteria.

Under the normal setting, it is shown that the SVM is two fifths to two thirds as effective as the LDA when the mean separation between two classes is substantially large with the Bayes error rate of 4 to 10%. Boosting is shown to be one fifth to one half as effective as the LDA in the same situation. Generally, the relative efficiency of algorithmic approach to modeling increases in the direction of growing overlap between classes and diminishes quickly as the two classes become sufficiently apart. However, we find that certain convex loss criteria work favorably for the normal setting. For instance, a smooth variant of the SVM with squared hinge loss is shown to be even more efficient than logistic regression.

To broaden the scope of the comparison, we also examine the first-order difference between the Bayes risk and the limiting minimal error of the classifiers under consideration, when the underlying model (or the class of discriminant functions) is incorrectly specified and thus the Bayes risk consistency is not guaranteed. In addition, we carry out a simulation study under the settings not covered by the theoretical analysis to touch on the issue of robustness to mislabeling error in the data.

The remainder of this paper is organized as follows. Section 2 describes the theoretical framework for comparisons of classification methods. Section 3 states general result about the limiting distribution of discriminant coefficients as M-estimators under some regularity conditions and its applications for various classification methods. As the main result, the ARE comparisons based on the limiting distributions are presented in Section 4 along with finite-sample comparisons of excess error rates. Section 5 provides further comparisons of the methods in terms of robustness to model-misspecification or data contamination. Section 6 contains numerical comparisons of some of the methods in a setting with growing input dimensions. Concluding remarks are in Section 7.

2 Background and Framework for Comparison

Consider a classification problem where multivariate attributes are measured for each subject and for a number of subjects, their class memberships are observed. Let $\mathbf{X} = (X_1, \dots, X_p) \in \mathcal{X} = \mathbb{R}^p$ denote the attributes or predictors and Y be the class label which takes one of, say, k nominal values, $\mathcal{Y} = \{1, \dots, k\}$. Training data consist of a set of n observation pairs, $\mathcal{D}_n = \{(\mathbf{x}_i, y_i), i = 1, \dots, n\}$, where $(\mathbf{x}_i, y_i)$'s are viewed as independent and identically distributed random outcomes of $(\mathbf{X}, Y)$ from some unknown distribution $P_{\mathbf{X},Y}$. Given the training data, we want to find a classification rule, $\phi : \mathcal{X} \to \mathcal{Y}$, which can be generalized to future cases from the same distribution $P_{\mathbf{X},Y}$ with a small error rate. Typically, the error of the rule ϕ over an individual case $(\mathbf{x}, y)$ is measured by the 0-1 loss $\rho(y, \phi(\mathbf{x})) = I(y \neq \phi(\mathbf{x}))$, and its overall error rate is given by the probability of error $R(\phi) = P(Y \neq \phi(\mathbf{X}))$. Then the theoretically optimal rule with the minimum error rate, which is also known as the Bayes decision rule ϕ_B, can be characterized as $\phi_B(\mathbf{x}) = \arg\max_{j \in \mathcal{Y}} P(Y = j \mid \mathbf{X} = \mathbf{x})$.

For simplicity, we focus on classification with binary outcomes only ($k = 2$) in this paper and use symmetric class labels $\mathcal{Y} = \{-1, 1\}$ whenever convenient. With the symmetric labels, the optimal rule is expressed succinctly as $\phi_B(\mathbf{x}) = \text{sgn}(p(\mathbf{x}) - 1/2)$, where $p(\mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x})$. Many classification procedures in consideration can be viewed as a way of obtaining a real-valued *discriminant* function $f : \mathbb{R}^p \to \mathbb{R}$, which induces a rule $\phi_f(\mathbf{x}) = \text{sgn}(f(\mathbf{x}))$, by minimizing the risk under a convex surrogate loss of the 0-1 loss. Generally, a class of functions $\mathcal{F}$ is specified a priori for the discriminant function f, for example, a linear space spanned by a set of basis functions or a reproducing kernel Hilbert space with a kernel function K. As we mainly consider the setting where ϕ_B is linear in $\mathbf{x}$, we will restrict $\mathcal{F}$ to linear discriminant functions only in this paper.

2.1 Normal Distribution Setting

Suppose that the attributes $\mathbf{X}$ arise from one of two p-dimensional normal populations with different means but the same covariance:

$$\begin{aligned} \mathbf{X} &\sim N_p(\boldsymbol{\mu}_+, \Sigma) \quad \text{with probability } \pi_+ \text{ and} \\ \mathbf{X} &\sim N_p(\boldsymbol{\mu}_-, \Sigma) \quad \text{with probability } \pi_-, \end{aligned} \tag{1}$$

where $\pi_+ = P(Y = 1)$, $\pi_- = P(Y = -1)$, and $\pi_+ + \pi_- = 1$. Fisher's linear discriminant analysis (LDA) is a standard example of linear classifiers in statistics, which is proven to be optimal in minimizing the misclassification rate under the normality and equal covariance assumptions. The optimal classification boundary is determined by Fisher's linear discriminant function $f^*(\mathbf{x}) = \beta_0^* + \boldsymbol{\beta}^{*\prime}\mathbf{x}$, where $\beta_0^* = \log(\pi_+/\pi_-) - (1/2)(\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-)'\Sigma^{-1}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)$ and $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)' = \Sigma^{-1}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)$.

Since this general LDA setting can be transformed to the canonical setting by means of a linear transformation, without loss of generality we will assume the following canonical LDA setting:

$$\begin{aligned} \mathbf{X}|\, Y = 1 &\sim N_p\left(\frac{\Delta}{2}\mathbf{e}_1, I\right) \quad \text{and} \\ \mathbf{X}|\, Y = -1 &\sim N_p\left(-\frac{\Delta}{2}\mathbf{e}_1, I\right), \end{aligned} \tag{2}$$

where $\mathbf{e}_1 = (1, 0, \ldots, 0)'$, I is the $p \times p$ identity matrix, and $\Delta = \{(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)'\Sigma^{-1}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)\}^{\frac{1}{2}}$ (known as the Mahalanobis distance between the two normal distributions). For the canonical setting, Fisher's linear discriminant coefficients are simplified to $\beta_0^* = \log(\pi_+/\pi_-)$ and $\boldsymbol{\beta}^* = \Delta\mathbf{e}_1$. The Bayes decision rule induced by f^* is then $\phi_B(\mathbf{x}) = \text{sgn}(f^*(\mathbf{x}))$.

2.2 Classification Methods

To make comparison of classification methods simple, we restrict the space of discriminant functions to linear functions of $\mathbf{x}$ only. Now consider deriving a linear discriminant function based on the training data. A host of classification methods can be applied to derive a linear discriminant rule.

If we model the data fully under the true LDA setting, we get the plug-in LDA discriminant function $\hat{f}_{LDA}$ with estimated means $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ and covariance Σ. As an intermediate method sitting half-way between the full modeling approach of the LDA and purely algorithmic approach, logistic regression models the conditional distribution of Y given $\mathbf{x}$ with the distribution of $\mathbf{x}$ unspecified. Also, it can be viewed as an M-estimator with deviance loss. The discriminant function of logistic regression $\hat{f}_{LR}$ is an estimate of the logit function,

$$f(\mathbf{x}) = \log(P(Y = 1|\mathbf{X} = \mathbf{x})/P(Y = -1|\mathbf{X} = \mathbf{x})) = \beta_0 + \boldsymbol{\beta}'\mathbf{x},$$

which is determined by maximizing the conditional log likelihood or minimizing the negative log likelihood of $(\beta_0, \boldsymbol{\beta}')' \in \mathbb{R}^{p+1}$:

$$L_n(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \log\left(1 + \exp(-y_i(\beta_0 + \boldsymbol{\beta}'\mathbf{x}_i))\right). \tag{3}$$

In contrast to the LDA and logistic regression, discriminant functions for algorithmic methods are obtained directly through consideration of the classification boundary itself, not the probability model underlying the data. For example, the linear support vector machine finds the optimal hyperplane $\beta_0 + \boldsymbol{\beta}'\mathbf{x} = 0$ with a large margin between two classes by minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(1 - y_i(\beta_0 + \boldsymbol{\beta}'\mathbf{x}_i))_+ + \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2, \tag{4}$$

where λ is a positive tuning parameter that controls the trade-off between the empirical risk under the hinge loss (the first term) and the inverse of the margin as a penalty (the second term). It attempts to minimize the error rate directly by using a convex surrogate loss of the misclassification count. The optimal hyperplane ($\hat{\beta}_0 + \hat{\boldsymbol{\beta}}'\mathbf{x} = 0$) found as a solution to (4) then yields the discriminant function for the SVM, $\hat{f}_{SVM}(\mathbf{x}) = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}'\mathbf{x}$. There are other variants of the SVM for large-margin classification as well. For instance, the smooth SVM (Lee and Mangasarian 2001) uses squared hinge loss as a loss criterion.

As another convex risk minimization method in machine learning, boosting (Freund and Schapire 1997) finds a discriminant function by sequentially updating the current fitted function with a weighted version of data and combining the sequence of the fitted functions. Although the discriminant function from boosting in general takes the form of a weighted sum of weak learners obtained stagewise, when linear functions are used as weak learners, the expanded function space for boosting stays the same as that for the weak learners because any convex combination of linear functions is still linear. For the reason, we take the simple view of boosting as an M-estimator minimizing

$$L_n(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \exp(-y_i(\beta_0 + \boldsymbol{\beta}'\mathbf{x}_i)) \tag{5}$$

in this paper, borrowing the perspective on boosting in Friedman et al. (2000).

Consequently, the loss criterion employed to determine the discriminant function characterizes the difference among logistic regression, the SVM, and boosting in terms of their statistical behavior and classification accuracy. Each of the three methods can be described as an M-estimator under the loss of binomial deviance $\rho(s) = \log(1+\exp(-s))$, hinge $(1-s)_+$, and exponential $\exp(-s)$, respectively, where $s \equiv yf(\mathbf{x})$ for $f(\mathbf{x}) = \beta_0 + \boldsymbol{\beta}'\mathbf{x}$.

The main focus of this paper is to theoretically examine the effect of bypassing probability modeling of data on the error rate. We investigate the issue by comparing the LDA, logistic regression, the SVM, and boosting, which represent a wide spectrum of classification procedures spanning from full model-based to algorithmic approaches.

2.3 Error Rates and Relative Efficiency

For a discriminant function $\hat{f}$ from training data, let $R(\hat{f}) \equiv R(\phi_{\hat{f}})$ be the error rate of the associated discriminant rule, $\phi_{\hat{f}}(\mathbf{x}) = \mathrm{sgn}(\hat{f}(\mathbf{x}))$. That is, for $(\mathbf{X}, Y)$ independent of the data used to determine $\hat{f}$, $R(\hat{f}) \equiv P(Y \neq \mathrm{sgn}(\hat{f}(\mathbf{X})) = P(Y\hat{f}(\mathbf{X}) < 0)$. Note that $R(\hat{f})$ is a random variable due to the fact that $\hat{f}$ depends on the training data. $R(\hat{f}) - R(\phi_B)$ represents the excess error rate of $\hat{f}$ compared to the Bayes decision rule ϕ_B with the minimum error rate. The Bayes error is given by

$$R(\phi_B) = \pi_+\Phi\left(-\frac{\Delta}{2} - \frac{1}{\Delta}\log\frac{\pi_+}{\pi_-}\right) + \pi_-\Phi\left(-\frac{\Delta}{2} + \frac{1}{\Delta}\log\frac{\pi_+}{\pi_-}\right),$$

where $\Phi(\cdot)$ is the standard normal cdf.

Efron (1975) compared logistic regression to the LDA in terms of the excess error rate by examining the asymptotic relative efficiency (ARE) of logistic regression (LR) to normal discrimination (LDA), which is defined as

$$\lim_{n\to\infty} \frac{E(R(\hat{f}_{LDA}) - R(\phi_B))}{E(R(\hat{f}_{LR}) - R(\phi_B))}.$$

In his analysis, logistic regression is shown to be between one half and two thirds as effective as normal discrimination typically. The key fact in the analysis is that for a linear discriminant method $\hat{f}(\mathbf{x}) = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}'\mathbf{x}$, if $\sqrt{n}(\underline{\hat{\boldsymbol{\beta}}} - \underline{\boldsymbol{\beta}}^*) \xrightarrow{d} N_{p+1}(\mathbf{0}, \Sigma_{\boldsymbol{\beta}})$ under the canonical setting, the expected excess error rate, $E(R(\hat{f}) - R(\phi_B))$ is given by

$$\frac{\pi_+\phi(D_1)}{2\Delta n}\left[\sigma_{00} - \frac{2\beta_0^*}{\Delta}\sigma_{01} + \frac{(\beta_0^*)^2}{\Delta^2}\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp}\right] + o\left(\frac{1}{n}\right), \tag{6}$$

where $\underline{\hat{\boldsymbol{\beta}}} = (\hat{\beta}_0, \hat{\boldsymbol{\beta}}')'$, $\underline{\boldsymbol{\beta}}^* = (\beta_0^*, \boldsymbol{\beta}^{*\prime})'$, $D_1 = \Delta/2 + (1/\Delta)\log(\pi_+/\pi_-)$, ϕ is the pdf of the standard normal distribution, and σ_{ij} is the ijth entry of $\Sigma_{\boldsymbol{\beta}}$ $(i, j = 0, \ldots, p)$. In other words, (6) shows that the mean increased error rate of $\hat{f}$ relative to ϕ_B can be expressed in terms of the variance of $\underline{\hat{\boldsymbol{\beta}}}$ in the limiting distribution. It indicates how the accuracy of the estimator $\underline{\hat{\boldsymbol{\beta}}}$ of $\underline{\boldsymbol{\beta}}^*$ affects the excess error rate of the discriminant rule with $\underline{\hat{\boldsymbol{\beta}}}$ as its coefficients. When two procedures are consistent in getting the Bayes decision rule, their asymptotic relative efficiency can be measured by the rate at which the expected excess error goes to zero as the sample size n grows. With the same parametric rate of $1/n$ in (6), we see that the efficiency of a procedure is determined by its leading coefficient of $1/n$ in the expression of the excess error. Borrowing Efron's theoretical framework, we extend the analysis to include those prediction-oriented modern classification tools.

3 Asymptotic Distribution of Discriminant Coefficients

To compare the relative efficiency of different ways to determine a linear discrimination rule, we need to identify the asymptotic distribution of the coefficient vector for each method first. For the LDA and logistic regression, Efron used large sample theory of the maximum likelihood estimators in exponential family distributions. Under the canonical setting, it was shown that the limit distribution of $\sqrt{n}(\underline{\hat{\boldsymbol{\beta}}} - \underline{\boldsymbol{\beta}}^*)$ for the LDA is $N_{p+1}(\mathbf{0}, \Sigma_{\boldsymbol{\beta}})$ with

$$\Sigma_{\boldsymbol{\beta}} = \frac{1}{\pi_+\pi_-}\begin{pmatrix} 1+\frac{\Delta^2}{4} & \frac{\Delta}{2}(\pi_+-\pi_-) & 0 & \cdots & 0 \\ \frac{\Delta}{2}(\pi_+-\pi_-) & 1+2\Delta^2\pi_+\pi_- & 0 & & \\ 0 & 0 & 1+\Delta^2\pi_+\pi_- & & \\ \vdots & & & \ddots & \\ 0 & & & & 1+\Delta^2\pi_+\pi_- \end{pmatrix}.$$

Since the coefficient vectors for logistic regression, the SVM and its variants, and boosting are defined as a minimizer of a convex loss criterion, asymptotic theories for M-estimators in van der Vaart (2000) and Hjort and Pollard (1993), for example, can be used to identify their limiting distributions. See also Pollard (1991), Geyer (1994), Knight and Fu (2000) and Rocha et al. (2009).

For general description of the asymptotics of M-estimators, let $L_n(\beta_0, \boldsymbol{\beta}) \equiv \sum_{i=1}^n \rho(y_i, \mathbf{x}_i; \beta_0, \boldsymbol{\beta})$ for a convex loss ρ (with respect to β_0 and $\boldsymbol{\beta}$). Using $\underline{\boldsymbol{\beta}}$ for short notation of $(\beta_0, \boldsymbol{\beta}')'$, define $\underline{\hat{\boldsymbol{\beta}}}$ as the minimizer of $L_n(\beta_0, \boldsymbol{\beta})$. Let $L(\underline{\boldsymbol{\beta}}) = E\rho(Y, \mathbf{X}; \underline{\boldsymbol{\beta}})$ be the true risk under ρ, and $\underline{\boldsymbol{\beta}}^*$ be the population risk minimizer, $arg\min L(\underline{\boldsymbol{\beta}})$.

Under the following regularity conditions (adapted from Rocha et al. (2009)) that

C1. $\underline{\boldsymbol{\beta}}^*$ is bounded and unique,

C2. $L(\underline{\boldsymbol{\beta}})$ is bounded for each $\underline{\boldsymbol{\beta}}$,

C3. $\rho(y, \mathbf{x}; \underline{\boldsymbol{\beta}})$ is differentiable with respect to $\underline{\boldsymbol{\beta}}$ at $\underline{\boldsymbol{\beta}} = \underline{\boldsymbol{\beta}}^*$ for almost every $(\mathbf{x}, y)$ in $P_{\mathbf{X},Y}$ with derivative $\dfrac{\partial\rho(y, \mathbf{x}; \underline{\boldsymbol{\beta}})}{\partial\underline{\boldsymbol{\beta}}}$ and $G(\underline{\boldsymbol{\beta}}^*) \equiv E\left(\dfrac{\partial\rho(Y, \mathbf{X}; \underline{\boldsymbol{\beta}}^*)}{\partial\underline{\boldsymbol{\beta}}}\right)\left(\dfrac{\partial\rho(Y, \mathbf{X}; \underline{\boldsymbol{\beta}}^*)}{\partial\underline{\boldsymbol{\beta}}}\right)'$,

C4. $L(\underline{\boldsymbol{\beta}})$ is twice differentiable with respect to $\underline{\boldsymbol{\beta}}$ at $\underline{\boldsymbol{\beta}} = \underline{\boldsymbol{\beta}}^*$ with positive definite Hessian matrix

$$H(\underline{\boldsymbol{\beta}}^*) \equiv \left.\frac{\partial^2 L(\underline{\boldsymbol{\beta}})}{\partial\underline{\boldsymbol{\beta}}\cdot\partial\underline{\boldsymbol{\beta}}'}\right|_{\underline{\boldsymbol{\beta}}=\underline{\boldsymbol{\beta}}^*},$$

we can establish asymptotic normality of $\underline{\hat{\boldsymbol{\beta}}}$ as an M-estimator and its consistency. The convexity of the loss ρ is a key condition in establishing the asymptotic normality of $\underline{\hat{\boldsymbol{\beta}}}$. Although it can be replaced with any set of conditions yielding uniform convergence of the risk functions over compact sets, the convexity condition would suffice for our discussion.

Theorem 1. *Under the regularity conditions C1-C4 with a convex loss ρ, the asymptotic distribution of $\underline{\hat{\boldsymbol{\beta}}} = arg\min_{\underline{\boldsymbol{\beta}}} \sum_{i=1}^n \rho(y_i, \mathbf{x}_i; \underline{\boldsymbol{\beta}})$ based on a random sample from a distribution $P_{\mathbf{X},Y}$ is*

$$\sqrt{n}(\underline{\hat{\boldsymbol{\beta}}} - \underline{\boldsymbol{\beta}}^*) \xrightarrow{d} N_{p+1}(\mathbf{0}, H(\underline{\boldsymbol{\beta}}^*)^{-1}G(\underline{\boldsymbol{\beta}}^*)H(\underline{\boldsymbol{\beta}}^*)^{-1}).$$

Note that the population minimizer $\underline{\boldsymbol{\beta}}^*$ depends on ρ, and under the canonical setting, $\underline{\boldsymbol{\beta}}^*$ may have a different scale than the optimal coefficients in the theoretical LDA, depending on the method used. The difference is to be discussed shortly.

3.1 Support Vector Machine

Koo et al. (2008) examined the limiting distribution of the linear SVM in general setting. Technically, the analysis exploits a close link between the SVM and median regression yet with categorical responses, and applies the results on absolute deviation regression estimators in Pollard (1991) to the linear SVM. Due to the penalty in (4) and a slightly different set of regularity conditions considered, the result in Koo et al. (2008) is not a direct application of Theorem 1. However, in a nutshell, it shows that when the effect of the penalty gradually diminishes with $\lambda = o(n^{-1/2})$, the penalized coefficients of the linear SVM behave in the same way as what Theorem 1 would predict asymptotically. In particular, it was shown in the paper that under the LDA setting with equal proportions for the two classes, the classification boundary of the linear SVM coincides with that of the LDA in the limit, ensuring classification consistency.

To extend the result to general case of arbitrary class proportions of π_+ and π_-, we first identify the optimal discriminant coefficients $\underline{\boldsymbol{\beta}}^*$ under the hinge loss by minimizing the risk $L(\underline{\boldsymbol{\beta}}) = E(1 - Y \cdot \underline{\boldsymbol{\beta}}'\mathbf{X})_+$ or equivalently finding the root of the equation $S(\underline{\boldsymbol{\beta}}) \equiv \frac{\partial}{\partial \underline{\boldsymbol{\beta}}} L(\underline{\boldsymbol{\beta}}) = \mathbf{0}$. Here we show the derivation of $\underline{\boldsymbol{\beta}}^*$ for the SVM in detail for illustration and take similar steps for other methods. Under the LDA setting, the equation $S(\underline{\boldsymbol{\beta}}) = \mathbf{0}$ becomes

$$\begin{aligned} \pi_+\Phi(a_p) &= \pi_-\Phi(a_m) & (7)\\ \left[\pi_+\phi(a_p) + \pi_-\phi(a_m)\right] \boldsymbol{\Sigma}^{\frac{1}{2}}\omega^* &= \pi_+\Phi(a_p)\boldsymbol{\mu}_+ - \pi_-\Phi(a_m)\boldsymbol{\mu}_-, & (8)\end{aligned}$$

where

$$a_p \equiv \frac{1 - \beta_0^* - \boldsymbol{\mu}_+'\boldsymbol{\beta}^*}{\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}^*\|}, \quad a_m \equiv \frac{1 + \beta_0^* + \boldsymbol{\mu}_-'\boldsymbol{\beta}^*}{\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}^*\|}, \quad \text{and } \omega^* \equiv \frac{\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}^*}{\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}^*\|}. \tag{9}$$

$\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of standard normal distribution, respectively.

Plugging (7) into (8) and solving for ω^*, we have

$$\omega^* = \frac{\pi_-\Phi(a_m)}{\pi_+\phi(a_p) + \pi_-\phi(a_m)} \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-). \tag{10}$$

Since the norm of ω^* must be 1, taking the norm of both sides of the above equation yields

$$\frac{\pi_-\Phi(a_m)}{\pi_+\phi(a_p) + \pi_-\phi(a_m)}\Delta = 1.$$

Further simplifying the equation using (7), we arrive at the following relation for a_p and a_m:

$$\frac{\phi(a_p)}{\Phi(a_p)} + \frac{\phi(a_m)}{\Phi(a_m)} = \Delta. \tag{11}$$

Then we can solve the equations (7) and (11) for a_p and a_m numerically. Once a_p and a_m are obtained, from the relation $a_p + a_m = \{2 - (\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)'\boldsymbol{\beta}^*\}/\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}^*\|$ and (10), we can get the optimal coefficients

$$\boldsymbol{\beta}^* = \frac{2}{[a_p + a_m + \Delta]\Delta}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-), \tag{12}$$

and the intercept

$$\beta_0^* = \frac{a_m - a_p}{a_p + a_m + \Delta} - \frac{1}{2}(\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-)'\boldsymbol{\beta}^*. \tag{13}$$

Clearly from (7), if $\pi_+ = \pi_-$, then $a_p = a_m$ (call it a^*), and a^* solves $\phi(a^*)/\Phi(a^*) = \Delta/2$. The values of a^* can be tabulated when Δ varies. See Table 1 for some a^* values given a range of Δ. In the balanced case, the optimal parameters become

$$\boldsymbol{\beta}^* = \frac{2}{(2a^* + \Delta)\Delta}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-), \quad \text{and} \quad \beta_0^* = -\frac{1}{2}(\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-)'\boldsymbol{\beta}^*.$$

Under the canonical LDA setting in particular, they are further simplified to

$$\boldsymbol{\beta}^* = \frac{2}{(2a^* + \Delta)}\mathbf{e}_1, \quad \text{and} \quad \beta_0^* = 0.$$

Note that the optimal parameters for linear SVM have the scale factor of $c_{\text{SVM}} \equiv 2/(2a^*\Delta + \Delta^2)$ when compared with the counterparts in Fisher's linear discriminant function, that is, $\underline{\boldsymbol{\beta}}^*_{\text{SVM}} = c_{\text{SVM}} \cdot \underline{\boldsymbol{\beta}}^*_{\text{LDA}}$.

From the results in Koo et al. (2008) with equal probabilities, we get

$$H(\underline{\boldsymbol{\beta}}^*) = \frac{\phi(a^*)(2a^* + \Delta)}{2}\begin{pmatrix} 1 & 0 & \cdots & & 0 \\ 0 & \frac{1}{4}(\Delta + 2a^*)^2 & & & \\ \vdots & & 1 & & \\ & & & \ddots & \\ 0 & & & & 1 \end{pmatrix}$$

and

$$G(\underline{\boldsymbol{\beta}}^*) = \Phi(a^*)\begin{pmatrix} 1 & 0 & \cdots & & 0 \\ 0 & -\frac{1}{4}(\Delta^2 + a^*\Delta - 4) & & & \\ \vdots & & 1 & & \\ & & & \ddots & \\ 0 & & & & 1 \end{pmatrix}.$$

Hence the variance matrix in the limiting distribution of $\hat{\underline{\boldsymbol{\beta}}}$ is given by

$$H(\underline{\boldsymbol{\beta}}^*)^{-1}G(\underline{\boldsymbol{\beta}}^*)H(\underline{\boldsymbol{\beta}}^*)^{-1} = \frac{4}{\Phi(a^*)}c^2_{\text{SVM}}\begin{pmatrix} 1 & 0 & \cdots & & 0 \\ 0 & \frac{-4(\Delta^2 + a^*\Delta - 4)}{(\Delta + 2a^*)^4} & & & \\ \vdots & & 1 & & \\ & & & \ddots & \\ 0 & & & & 1 \end{pmatrix}.$$

From the expression of the excess error in (6) and consideration of the scale factor c_{SVM}, we can verify that the asymptotic relative efficiency of the linear SVM to LDA is given by

$$\text{ARE}_{SVM} = \Phi(a^*)(1+\frac{\Delta^2}{4}) = \frac{2\phi(a^*)}{\Delta}(1+\frac{\Delta^2}{4}),$$

where a^* is the constant satisfying $\phi(a^*)/\Phi(a^*) = \Delta/2$. Some values of the relative efficiency corresponding to a range of class separation Δ will be given later for more concrete comparisons.

It is important to observe that when $\pi_+ \neq \pi_-$ in general, only the optimal coefficients $\boldsymbol{\beta}^*$ (without the intercept) of linear SVM in (12) are proportional to those of LDA, which results in inconsistency. Remedies for the inconsistency would require alternative ways to estimate the intercept given $\hat{\boldsymbol{\beta}}$.

3.2 Variants of Support Vector Machine

There are variants of the SVM built around smooth versions of the hinge loss motivated mainly for computational ease. However, changes in the loss criterion lead to different asymptotic behavior of the resulting discriminant functions.

3.2.1 Smooth SVM

Smooth SVM (Lee and Mangasarian 2001) refers to a variant of the SVM where the hinge loss criterion is replaced with its square version $\rho(s) = [(1-s)_+]^2$. The discriminant coefficients $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ are found by minimizing

$$\frac{1}{n}\sum_{i=1}^{n}[(1-y_i(\beta_0+\boldsymbol{\beta}'\mathbf{x}_i))_+]^2 + \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2. \tag{14}$$

In contrast to the hinge loss, the squared hinge loss is differentiable everywhere.

Following similar steps taken for the SVM, we can identify the optimal parameters $\underline{\boldsymbol{\beta}}^*$ under squared hinge loss as

$$\beta_0^* = \frac{a_m - a_p}{a_p + a_m + \Delta} - \frac{1}{2}(\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-)'\boldsymbol{\beta}^* \text{ and } \boldsymbol{\beta}^* = \frac{2}{[a_p + a_m + \Delta]\Delta}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-),$$

where a_p and a_m are the constants that solve the following equations with $\Theta(z) \equiv z\Phi(z) + \phi(z)$:

$$\begin{aligned} \pi_+\Theta(a_p) &= \pi_-\Theta(a_m) & (15)\\ \pi_+\Theta(a_p)\Delta &= \pi_+\Phi(a_p) + \pi_-\Phi(a_m). & (16) \end{aligned}$$

As in the standard SVM, the optimal parameters for the smooth SVM are not proportional to those for the LDA in general if the probabilities of the two classes are not equal. If $\pi_+ = \pi_-$, the equation (15) becomes $\Theta(a_p) = \Theta(a_m)$. Since $\Theta(z) = z\Phi(z) + \phi(z)$ is an increasing function of z, we conclude $a_p = a_m$. From the equation (16), the common value

a^* is given as the constant that solves $[a\Phi(a) + \phi(a)]\Delta = 2\Phi(a)$. The third column in Table 1 shows the values of a^* for the smooth SVM corresponding to the given values of Δ.

Under the canonical LDA setting with equal probabilities, the optimal parameters reduce to $\beta_0^* = 0$ and $\boldsymbol{\beta}^* = \dfrac{2}{(2a^* + \Delta)}\mathbf{e}_1$. Thus $\underline{\boldsymbol{\beta}}^*_{\text{SSVM}} = c_{\text{SSVM}} \cdot \underline{\boldsymbol{\beta}}^*_{\text{LDA}}$ with $c_{\text{SSVM}} \equiv 2/(2a^*\Delta + \Delta^2)$. When $\lambda = o(n^{-1/2})$, the smooth SVM in (14) provides a consistent estimator of $c_{\text{SSVM}} \cdot \underline{\boldsymbol{\beta}}^*_{\text{LDA}}$. Further with the Hessian matrix $H(\underline{\boldsymbol{\beta}}^*)$ and $G(\underline{\boldsymbol{\beta}}^*)$ for squared hinge loss, we can find the limiting covariance matrix and verify that the ARE of the smooth SVM to LDA is given by

$$\text{ARE}_{SSVM} = \frac{4\Phi(a^*)(1 + \Delta^2/4)}{\Delta(2a^* + \Delta)} = 2c_{\text{SSVM}}\Phi(a^*)(1 + \Delta^2/4).$$

Details of the derivation of $\underline{\boldsymbol{\beta}}^*$ and the asymptotic distribution of discriminant coefficients for the smooth SVM are given in Appendix.

Table 1: Values of a^* as a function of Δ for the linear SVM and its variants

Δ	SVM	Smooth SVM	Huberized SVM		
			$k = -1.5$	$k = -1$	$k = 0$
1.0	0.518	1.937	1.937	1.934	1.371
1.5	0.076	1.071	1.071	1.062	0.691
2.0	−0.303	0.481	0.480	0.467	0.184
2.5	−0.647	0.006	0.003	−0.011	−0.242
3.0	−0.969	−0.407	−0.411	−0.426	−0.621
3.5	−1.276	−0.782	−0.786	−0.802	−0.971
4.0	−1.572	−1.131	−1.136	−1.151	−1.301

3.2.2 Huberized SVM

The Huberized SVM (Rosset and Zhu 2007) is another variant of the SVM inspired by Huber's loss for robust regression. It retains the robustness of the SVM for large margin classification yet with differentiability in the loss. It replaces the hinge loss in (4) with

$$\rho_k(y, \mathbf{x}; \beta_0, \boldsymbol{\beta}) = \begin{cases} 2(k-1)y(\beta_0 + \boldsymbol{\beta}'\mathbf{x}) + (1 - k^2) & \text{if } y(\beta_0 + \boldsymbol{\beta}'\mathbf{x}) < k, \\ [1 - y(\beta_0 + \boldsymbol{\beta}'\mathbf{x})]^2 & \text{if } k \le y(\beta_0 + \boldsymbol{\beta}'\mathbf{x}) < 1, \\ 0 & \text{if } y(\beta_0 + \boldsymbol{\beta}'\mathbf{x}) \ge 1, \end{cases}$$

where $k < 1$ and as a bending constant, it demarcates the quadratic part of the loss. When k tends to $-\infty$, ρ_k approaches the squared hinge loss in the smooth SVM.

The expressions of the optimal parameters under Huberized hinge loss are shown to be the same as those of the smooth SVM:

$$\beta_0^* = \frac{a_m - a_p}{a_p + a_m + \Delta} - \frac{1}{2}(\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-)'\boldsymbol{\beta}^* \text{ and } \boldsymbol{\beta}^* = \frac{2}{[a_p + a_m + \Delta]\Delta}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-),$$

except that a_p and a_m are now defined as the constants that solve slightly different equations:

$$\begin{aligned} \pi_+[\Theta(a_p) - \Theta(a_p^k)] &= \pi_-[\Theta(a_m) - \Theta(a_m^k)] \qquad (17) \\ \pi_+[\Theta(a_p) - \Theta(a_p^k)]\Delta &= \pi_+[\Phi(a_p) - \Phi(a_p^k)] + \pi_-[\Phi(a_m) - \Phi(a_m^k)]. \qquad (18) \end{aligned}$$

Here a_p^k and a_m^k are the additional constants depending on k that are defined as

$$a_p^k = a_p - \frac{1-k}{\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}^*\|}, \quad a_m^k = a_m - \frac{1-k}{\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}^*\|}, \quad \text{and} \quad \|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}^*\| = \frac{2}{a_p + a_m + \Delta}.$$

Similar to other SVM type methods, except for the balanced case of $\pi_+ = \pi_-$, only the optimal parameters $\boldsymbol{\beta}^*$ of the Huberized SVM are parallel to those of the LDA in general. For the balanced case, we can show that $a_p = a_m (\equiv a^*)$, and $a_p^k = a_m^k (\equiv a_k^*) = ka^* - (1-k)\Delta/2$ from (17), and the identities further simplify (18) to

$$[\Theta(a^*) - \Theta(ka^* - (1-k)\Delta/2)]\Delta = 2[\Phi(a^*) - \Phi(ka^* - (1-k)\Delta/2)]. \qquad (19)$$

Given Δ and fixed k, we can solve the equation for a^*. The last three columns in Table 1 show the values of a^* corresponding to the given values of Δ when $k = -1.5$, -1, and 0, respectively.

In particular, under the balanced canonical LDA setting, the optimal parameters of the Huberized SVM are given by $\beta_0^* = 0$ and $\boldsymbol{\beta}^* = \frac{2}{(2a^* + \Delta)}\mathbf{e}_1$, where a^* is the constant satisfying (19). This yields $\underline{\boldsymbol{\beta}}^*_{\text{HSVM}} = c_{\text{HSVM}} \cdot \underline{\boldsymbol{\beta}}^*_{\text{LDA}}$ with $c_{\text{HSVM}} \equiv 2/(2a^*\Delta + \Delta^2)$ taking the same form as in the smooth SVM. With the limiting covariance matrix given in Appendix, we can show that the ARE of the Huberized SVM to LDA is given by

$$\text{ARE}_{HSVM} = 2c_{\text{HSVM}}\big(1 + \frac{\Delta^2}{4}\big)\frac{[\Phi(a^*) - \Phi(a_k^*)]^2}{\Phi(a^*) - \Phi(a_k^*) + \frac{\Delta}{2}(k-1)\Theta(a_k^*)}.$$

3.3 Boosting

Similarly, the limiting distribution of the discriminant coefficients for boosting can be found, and its efficiency relative to the LDA in terms of the excess error rate can be evaluated. Without diminishing the merit of boosting in expanding a model space with weak learners in the LDA setting, we simply define the boosting estimator $\underline{\hat{\boldsymbol{\beta}}}$ as the minimizer of (5) under the exponential loss criterion $\rho(y, \mathbf{x}; \beta_0, \boldsymbol{\beta}) = \exp(-y(\beta_0 + \boldsymbol{\beta}'\mathbf{x}))$.

In the general LDA setting, the optimal discriminant coefficients under the exponential loss is given as

$$\beta_0^* = \frac{1}{2}\left(\log\frac{\pi_+}{\pi_-} - \frac{1}{2}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-)\right) \text{ and } \boldsymbol{\beta}^* = \frac{1}{2}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-).$$

First, we see that for every π_+ and π_-, the optimal coefficient vector $\underline{\boldsymbol{\beta}}^*$ for boosting is proportional to that of LDA, with the proportional constant $c_{\text{boost}} = 1/2$. Thus $\underline{\hat{\boldsymbol{\beta}}}$ is a consistent estimator of $(1/2)\underline{\boldsymbol{\beta}}^*_{\text{LDA}}$ in general. This ensures the Bayes risk consistency of boosting. See

Appendix for the derivation of $\underline{\boldsymbol{\beta}}^*$ and the asymptotic distribution of discriminant coefficients.

Under the canonical LDA setting (2) with equal class proportions, the asymptotic relative efficiency of boosting to LDA is given by

$$\mathrm{ARE}_{boost} = \frac{1 + \Delta^2/4}{\exp(\Delta^2/4)}.$$

Notice that the denominator is an exponential function of Δ, which implies that the relative efficiency of boosting drops very quickly as Δ grows. This is attributed to the characteristic of boosting that it tends to focus heavily on misclassified cases or outliers in contrast to the LDA, which is based on the average pattern of the cases.

4 Comparisons under Normal Setting

We compare the classification procedures theoretically by evaluating their relative efficiency for various degrees of class separation Δ. To contrast model-based classification methods with algorithmic methods in a simple theoretical framework, we focus mainly on the balanced case of $\pi_+ = \pi_-$ in this section, where all the methods in consideration are proven to be consistent.

4.1 Theoretical Comparison with ARE

To cover scenarios with different degree of overlap between the two classes, we vary Δ from 1 to 4. This range of Δ corresponds to the Bayes error rates from 31% to 2% approximately. Table 2 gives the ARE values of the various methods considered in the foregoing section. First of all, the ARE values are less than one for each method compared to the LDA as a plug-in rule with the maximum likelihood estimators of the model parameters. So, the main focus of comparison in this normal setting is how much efficiency is lost in reducing the error rate when we bypass modeling of the underlying probability distribution in full.

The SVM and boosting, as large margin classifiers widely used in applications, are shown to be less efficient than logistic regression across the range of Δ values. In other words, modeling at least the conditional probability helps use data more efficiently than maximizing classification margins under the hinge loss or the exponential loss. The efficiency of both methods relative to LDA diminishes quickly when the two classes become more separable. Between the two, the SVM is slightly more efficient than boosting as the latter heavily focuses on outlying observations near the classification boundary. Especially when two classes are nearly separable, boosting becomes very ineffective in using data. When the Bayes error is less than 6%, boosting requires more than twice the data needed for the SVM to attain the same accuracy asymptotically.

Among the SVM and its variants, surprisingly, the smooth SVM turns out to be very efficient, even better than logistic regression and the vanilla SVM for each value of Δ. Probably, it can be, in part, explained by the analogue that regression with the squared error loss is often more efficient than its counterpart with the absolute deviation loss in estimating the mean especially in such situations as the normal setting. More explicitly, we note that there is a close connection between Fisher's LDA and a naive regression approach to classification with class labels as the response. It can be shown that the least squares coefficient $\hat{\boldsymbol{\beta}}$ is identical up to a scalar multiple to the LDA coefficient, that is, $\hat{\boldsymbol{\beta}} \propto \hat{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}_+ - \hat{\boldsymbol{\mu}}_-)$; see, for example, an exercise in Chapter 4 of Hastie et al. (2001). From the relation

$$\sum_{i=1}^{n}(y_i - \beta_0 - \boldsymbol{\beta}'\mathbf{x}_i)^2 = \sum_{i=1}^{n}(1 - y_i(\beta_0 + \boldsymbol{\beta}'\mathbf{x}_i))^2,$$

we see that squaring the hinge loss has a similar effect as the squared error loss. Moreover, the asymptotic analysis with the squared error loss confirms that the naive regression approach to classification is equivalent to the LDA in the limit under the canonical LDA setting with equal proportions. Raising the power of the hinge loss further to three did not improve the smooth SVM in terms of ARE values. The cubed hinge loss was comparable to logistic regression (the results not shown in the table).

As an intermediate method, the Huberized SVM lies generally between the SVM and the smooth SVM in terms of the relative efficiency. As the bending constant k decreases, the Huberized SVM approaches the smooth SVM and its relative efficiency converges to that of the smooth SVM. When k is as small as -1.5 as shown in Table 2, the Huberized SVM is virtually as efficient as the smooth SVM.

Table 2: Asymptotic relative efficiency of classification methods to LDA under the canonical LDA setting in (2) with equal proportions

Δ	Bayes Error	Logistic Regression	SVM	Boosting	Smooth SVM	Huberized SVM		
						$k=-1.5$	$k=-1$	$k=0$
1.0	0.3085	0.995	0.872	0.974	0.999	0.999	0.999	0.968
1.5	0.2266	0.968	0.829	0.890	0.981	0.981	0.981	0.939
2.0	0.1587	0.899	0.762	0.736	0.925	0.925	0.924	0.876
2.5	0.1056	0.786	0.664	0.537	0.820	0.820	0.818	0.771
3.0	0.0668	0.641	0.541	0.343	0.678	0.678	0.676	0.633
3.5	0.0401	0.486	0.411	0.190	0.521	0.520	0.518	0.483
4.0	0.0228	0.343	0.290	0.092	0.371	0.371	0.369	0.343

NOTE: The column for logistic regression has been taken from Efron (1975).

When the class proportions are not equal, the SVM and its variants are not consistent as discussed in the previous section. Figure 1 shows how the excess errors of the infinite-sample version of SVM, Huberized SVM and smooth SVM vary as π_+ increases from 0.5 to 0.92 when $\Delta = 2$ in the canonical LDA setting. The Bayes error rate decreases from 15.9% to 5.9% as π_+ increases as indicated by the dotted line. Note that the axis for the Bayes error is on the right. Clearly, as the extent of imbalance between two classes increases, the excess error gets larger for the three methods. However, the size of the increased error due to the bias in estimation of intercept by those methods appears very small compared to the Bayes error. Since their excess errors do not converge to zero in the limit, the asymptotic relative efficiency of each of the methods to LDA will be zero.

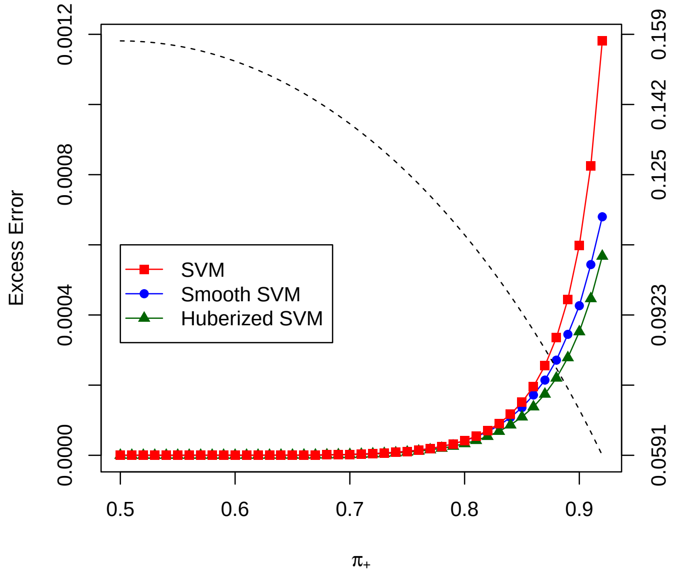

Figure 1: Excess errors of the infinite-sample version of SVM, Huberized SVM, and smooth SVM as the proportion of a positive class (π_+) varies from 0.5 to 0.92 in the canonical LDA setting with $\Delta = 2$. The dotted line indicates the Bayes error as a function of π_+ whose axis is on the right.

By contrast, logistic regression and boosting are consistent in general and they can be compared with LDA in terms of relative efficiency when there is imbalance between two classes. Using the result of the asymptotic distribution of discriminant coefficients for boosting in Appendix, we can show that under the canonical setting with general π_+, the relative efficiency of boosting to LDA is given by

$$\text{ARE}_{boost} = \frac{1 + \frac{\Delta^2}{4} - (\pi_+ - \pi_-)\log\frac{\pi_+}{\pi_-} + (2\pi_+\pi_- + \frac{1}{\Delta^2})(\log\frac{\pi_+}{\pi_-})^2 + (p-1)(1 + \Delta^2\pi_+\pi_-)}{\exp(\frac{\Delta^2}{4})\left(1 - (\pi_+ - \pi_-)\log\frac{\pi_+}{\pi_-} + (\frac{1}{4} + \frac{1}{\Delta^2})(\log\frac{\pi_+}{\pi_-})^2 + (p-1)\right)}.$$

Table 3 gives the relative efficiency values of boosting to LDA as π_+ increases from 0.5 to 0.9 when $\Delta = 2$ or 3. It is sufficient to consider the cases of $p = 1$ and $p \to \infty$ as in Efron's original comparison of logistic regression to LDA since the ARE for general p is given as a weighted average of the relative efficiencies of the two cases. The result indicates that with growing dimensionality, both methods become increasingly less efficient in comparison with LDA as the extent of imbalance between two classes increases. The relative merit of logistic regression to boosting seems to stay the same regardless of π_+ values.

Table 3: Asymptotic relative efficiency of logistic regression and boosting to LDA under the canonical LDA setting

Δ	π_+	Bayes Error	Logistic Regression		Boosting	
			$p=1$	$p\to\infty$	$p=1$	$p\to\infty$
2.0	0.5	0.1587	0.899	0.899	0.736	0.736
	0.6	0.1538	0.906	0.892	0.749	0.721
	0.667	0.1449	0.913	0.879	0.767	0.695
	0.75	0.1270	0.915	0.855	0.770	0.644
	0.9	0.0701	0.804	0.801	0.515	0.500
3.0	0.5	0.0668	0.641	0.641	0.343	0.343
	0.6	0.0651	0.649	0.633	0.352	0.333
	0.667	0.0619	0.662	0.618	0.368	0.316
	0.75	0.0554	0.682	0.589	0.391	0.283
	0.9	0.0337	0.667	0.511	0.310	0.191

4.2 Numerical Comparison with Finite-Sample Excess Error

To complement the theoretical comparison for large sample case, we carried out numerical comparisons of the expected excess error rates of the procedures for finite samples by varying the sample sizes from small to large (50 to 1000).

Given sample size n, we generated training data from the canonical LDA setting (2) with five covariates and equal class proportions ($\pi_+ = \pi_-$). Two values of Δ were considered for simulation: $\Delta = 2$ with the Bayes error of 15.87%, and $\Delta = 3$ with the Bayes error of 6.68%. Then we applied the classification methods in Table 2 to each simulated data set, and calculated the excess error of each method analytically using the following equation for the error of a linear classification rule, $\phi(\mathbf{x}) = \mathrm{sgn}(\beta_0 + \beta'\mathbf{x})$ under the canonical setting:

$$R(\phi) = \pi_+\Phi\bigg(-\frac{\beta_0 + \frac{\Delta}{2}\beta_1}{\|\beta\|}\bigg) + \pi_-\Phi\bigg(\frac{\beta_0 - \frac{\Delta}{2}\beta_1}{\|\beta\|}\bigg).$$

We repeated this process 1000 times for each sample size, and estimated the expected excess error rate by using the average of the data-specific error rates over the 1000 replicates. To reduce the variance of the mean excess error estimate due to different data realization along the sequence of sample sizes, we generated nested training data; training data with smaller sizes are always included in the training data with larger sizes.

Figure 2 shows the mean excess error rates of the classification methods estimated for finite samples. The scale on the x-axis is $1/n$. The estimated mean excess error curves show strong linear relationship with $1/n$ as the asymptotic theory suggests. Overall the excess error decreases to 0 at the rate of $1/n$. The LDA has the smallest slope (fastest reduction in error) followed by the smooth SVM, logistic regression, the SVM, and boosting in the respective order as implied by the ARE comparison. This result suggests that the asymptotic comparisons are relevant even in moderate sample size cases. Table 2 indicates that as two classes become more separable, the efficiency of the other classification methods

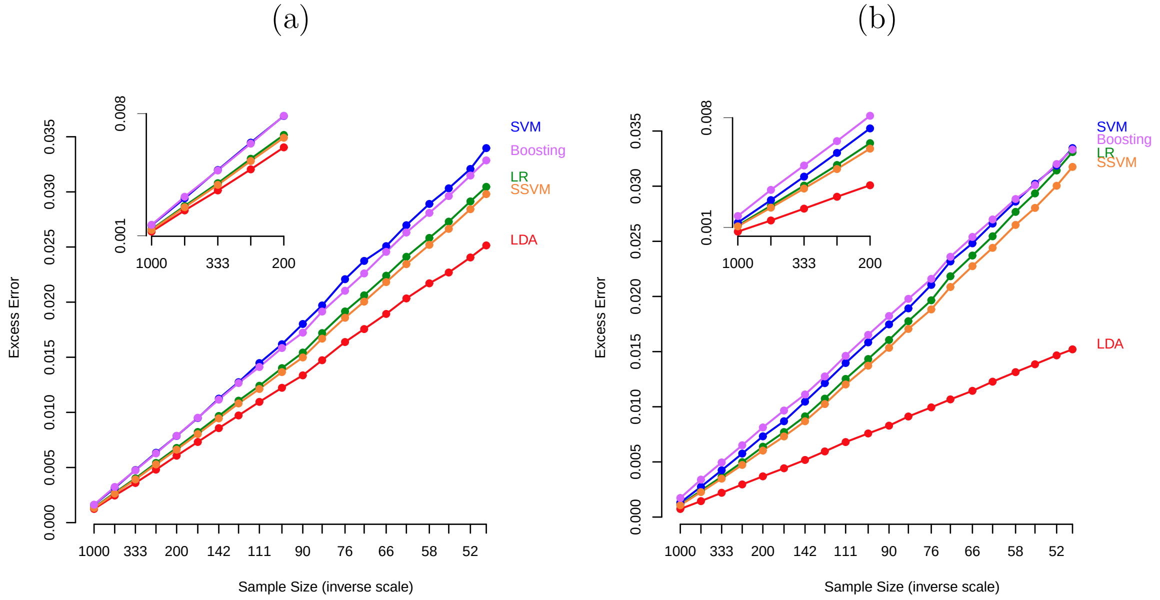

Figure 2: Finite-sample mean excess error rates of classification methods as the sample size varies. Data are simulated from the five dimensional canonical LDA setting with equal class proportions and (a) $\Delta = 2$ (Bayes error of 15.87%); (b) $\Delta = 3$ (Bayes error of 6.68%).

relative to the LDA drops. Figure 2 confirms that the increasing loss of efficiency occurs with the larger value of Δ for all the other methods in finite-sample cases.

5 Comparisons under Model Mis-specification or Data Contamination

The analysis so far is under the LDA setting, expectedly yielding favorable results for the LDA. In practice, there are many factors that may complicate proper modeling of data. For instance, a model could be misspecified or part of data may not follow the specified model even when it is correctly specified. By taking into account such realistic constraints in data modeling, we consider two scenarios different from the LDA setting for more comprehensive comparisons.

5.1 Mislabeling in LDA Setting

In the first scenario, we compare the robustness of the SVM and its variants when there is class label noise. To generate mislabeled data, we first simulate data from the canonical LDA setting and flip the class labels of a certain proportion of cases selected at random.

For simulation, the proportions of ±1 were set to equal, $p = 5$, and $\Delta = 2.7$, which yields the Bayes error rate of 8.851%. The sample size was $n = 100$, and 400 replicates of mislabeled samples with varying perturbation fractions were generated. Estimated discriminant functions were then evaluated in terms of the excess error rate from the Bayes error.

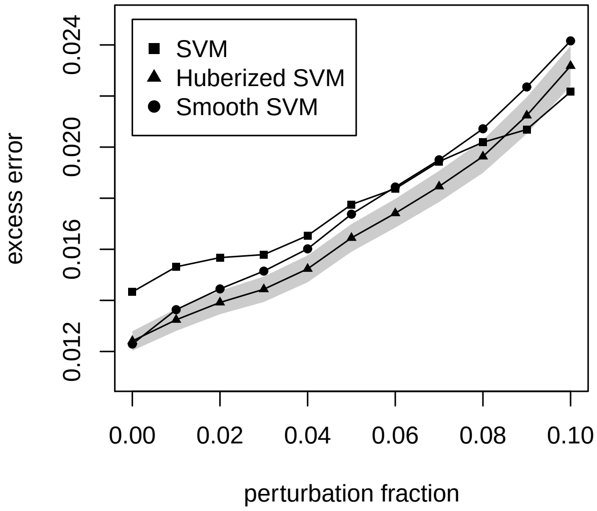

Figure 3: Mean excess error rates of SVM and its variants from 400 replicates as the mislabeling proportion varies when $\Delta = 2.7$, $p = 5$, and $\pi_+ = \pi_-$ with the Bayes error rate of 8.851%, and $n = 100$. The gray band indicates one standard error bound around the mean estimate for Huberized SVM from the replicates.

Figure 3 displays the mean excess error rates of the SVM, smooth SVM and Huberized SVM with $k = -0.5$ over the replicates as the perturbation fraction increases. When the mislabeling rate is small, the smooth SVM remains to be better than the SVM. However, as the mislabeling rate increases, the SVM results in lower error rate than the smooth SVM due to its robustness, analogous to the sample median and mean comparison in location parameter estimation. The Huberized SVM as a hybrid method strikes a balance between the two by combining the squared hinge loss and the plain hinge loss. The result indicates a trade-off between efficiency and robustness.

5.2 QDA Setting

The second scenario is when the Bayes error rate is not attainable due to model misspecification for model-based procedures and limitation in the family of discriminant functions for algorithmic procedures. As a scenario closely related to but different from the LDA setting, we consider a quadratic discriminant analysis (QDA) setting, where the covariance of one class is a scalar multiple of that of the other class:

$$\begin{aligned} \mathbf{X} &\sim N_p(\boldsymbol{\mu}_+, \boldsymbol{\Sigma}) \quad \text{with} \quad \text{probability} \quad \pi_+ = P(Y = 1) \quad \text{and} \\ \mathbf{X} &\sim N_p(\boldsymbol{\mu}_-, C\boldsymbol{\Sigma}) \quad \text{with} \quad \text{probability} \quad \pi_- = P(Y = -1) \end{aligned} \tag{20}$$

with a constant C greater than 1. Under this setting with $C > 1$, the Bayes boundary is no longer linear. Hence all linear classification methods compared here cannot be consistent, and we need to take into account the inconsistency of the procedures in comparison.

For more meaningful comparison, consider the following decomposition of the excess error of a rule $\phi_n \in \mathcal{F}$ (a restricted class of discriminant functions, for example, linear functions in our case) based on a sample of size n:

$$R(\phi_n) - R(\phi_B) = \{R(\phi_n) - R(\phi_\infty)\} + \{R(\phi_\infty) - R(\phi_{\mathcal{F}})\} + \{R(\phi_{\mathcal{F}}) - R(\phi_B)\},$$

where $\phi_{\mathcal{F}} = \arg\min_{\phi \in \mathcal{F}} R(\phi)$, and ϕ_∞ is the limiting rule of ϕ_n as n goes to ∞. The first error difference on the right hand side is called the *estimation error* in the machine learning literature. It is due to finite sample and converges to zero as n increases. The third difference is known as the *approximation error*, which is due to the restriction of $\mathcal{F}$ and common to all the linear procedures. It indicates the non-ignorable gap between the smallest error rate attainable within the class $\mathcal{F}$ and the Bayes error rate. Since the limiting rule ϕ_∞ depends on the method used to choose ϕ_n from $\mathcal{F}$, we call the first term a method-specific estimation error and the second term a method-specific approximation error. The method-specific approximation error is the key to capturing differences among the linear procedures in this QDA setting, providing the first order comparison.

For a linear classifier $\phi \in \mathcal{F}$ with discriminant coefficients $\underline{\boldsymbol{\beta}}$ in the QDA setting, the error rate of ϕ is given by

$$R(\phi) \equiv R(\underline{\boldsymbol{\beta}}) = \pi_+ \Phi\left(-\frac{\boldsymbol{\beta}'\boldsymbol{\mu}_+ + \beta_0}{\sigma}\right) + \pi_- \left[1 - \Phi\left(-\frac{\boldsymbol{\beta}'\boldsymbol{\mu}_- + \beta_0}{\sqrt{C}\sigma}\right)\right], \qquad (21)$$

where $\sigma \equiv \sqrt{\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}}$. Then the optimal linear classifier $\phi_{\mathcal{F}}$ can be identified with the minimizer $\underline{\boldsymbol{\beta}}^*$ of $R(\underline{\boldsymbol{\beta}})$:

$$\begin{aligned}
\boldsymbol{\beta}^* &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-), \\
\beta_0^* &= \sqrt{\frac{C\{(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)'\boldsymbol{\beta}^*\}^2}{(C-1)^2} + \frac{C\sigma^2(2\log\frac{\pi_+}{\pi_-} + \log C)}{C-1}} - \frac{C}{C-1}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)'\boldsymbol{\beta}^*.
\end{aligned}$$

To compute the method-specific approximation error of each method under consideration, we first obtain the limiting classification rule ϕ_∞ within $\mathcal{F}$ by applying large sample theory to the sample discriminant coefficients. Application of the standard asymptotics to the LDA and Theorem 1 to the rest of the methods yields the desired limiting rules. Due to page limitations, the results are omitted.

For numerical illustration, suppose that $p = 10$, $\pi_+ = 0.5$, $\boldsymbol{\Sigma} = \mathrm{I}$, $\boldsymbol{\mu}_+ = (-1, 0, \ldots, 0)'$, and $\boldsymbol{\mu}_- = (1, 0, \ldots, 0)'$ (hence with $\Delta \equiv \{(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)'\Sigma^{-1}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)\}^{\frac{1}{2}} = 2$) in the QDA setting of (20). The Bayes error rate is achieved by the theoretical quadratic discriminant analysis and can be expressed as $R(\phi_B) = \pi_+ P(\chi^2_{p,\lambda_1} > M) + \pi_- P(\chi^2_{p,\lambda_2} < \dfrac{M}{C})$, where $M = \dfrac{C}{(1-C)^2}\Delta^2 - \dfrac{C}{1-C}(2\log\dfrac{\pi_+}{\pi_-} + p\log C)$, and χ^2_{p,λ_1} and χ^2_{p,λ_2} are the chi-square random variables with degrees of freedom p and non-centrality parameters $\lambda_1 = \dfrac{\Delta^2}{(1-C)^2}$, and $\lambda_2 = \dfrac{C\Delta^2}{(1-C)^2}$, respectively.

Figure 4(a) depicts the decomposition of the minimum linear classification error $R(\phi_{\mathcal{F}})$ into the Bayes error $R(\phi_B)$ and the approximation error, $R(\phi_{\mathcal{F}}) - R(\phi_B)$, as we vary C from

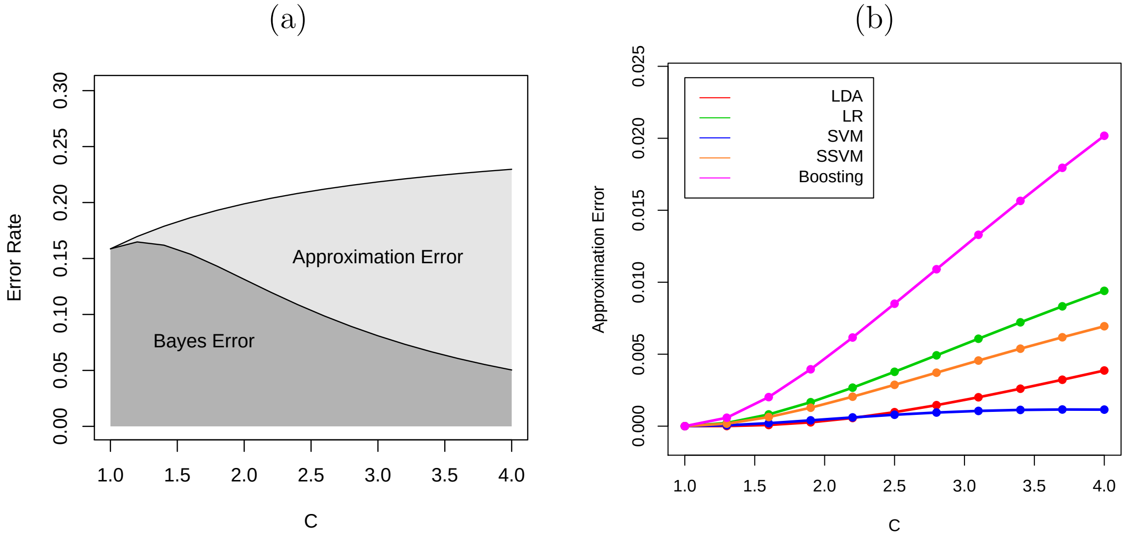

Figure 4: (a) Decomposition of the minimum linear classification error into the Bayes error and the approximation error; (b) method-specific approximation error of linear classifiers in the QDA setting with $\Delta = 2$, $\Sigma = I$, $p = 10$, and $\pi_+ = \pi_-$.

1 to 4. The height of the dark gray area shows the Bayes error rate, and that of the light gray region indicates the approximation error when $\mathcal{F}$ is restricted to linear classifiers in the QDA setting. Figure 4(b) shows the method-specific approximation errors for comparison as C varies. When $C > 1$, the limiting error rates of the methods are greater than the smallest linear classification error in general. However, the effect of "model misspecification" or restriction to linear classifiers when, in fact, quadratic discriminant functions are needed, differs among the methods in terms of the increased error, $R(\phi_\infty) - R(\phi_{\mathcal{F}})$. As the extent of misspecification increases, the linear SVM turns out to be most robust to the change of C, followed by the LDA and smooth SVM. The approximation error of boosting grows more substantially than the other methods as C increases.

6 Comparisons under Growing Input Dimensions

We examine the relative performance of model-based methods to algorithmic methods numerically in a setting where the input dimension grows. When the dimension is high and the sample size is small, it is paramount to regularize discriminant functions for control of variance and enhanced performance. In our numerical study, we mainly compare normal discriminant analysis, logistic regression and SVM by considering their penalized versions. For instance, the diagonal LDA (or naive LDA) is a constrained version of LDA with a diagonal covariance estimate. Just as the standard SVM penalizes the discriminant coefficients with their ℓ_2 norm, logistic regression can be modified with ℓ_2 norm penalty.

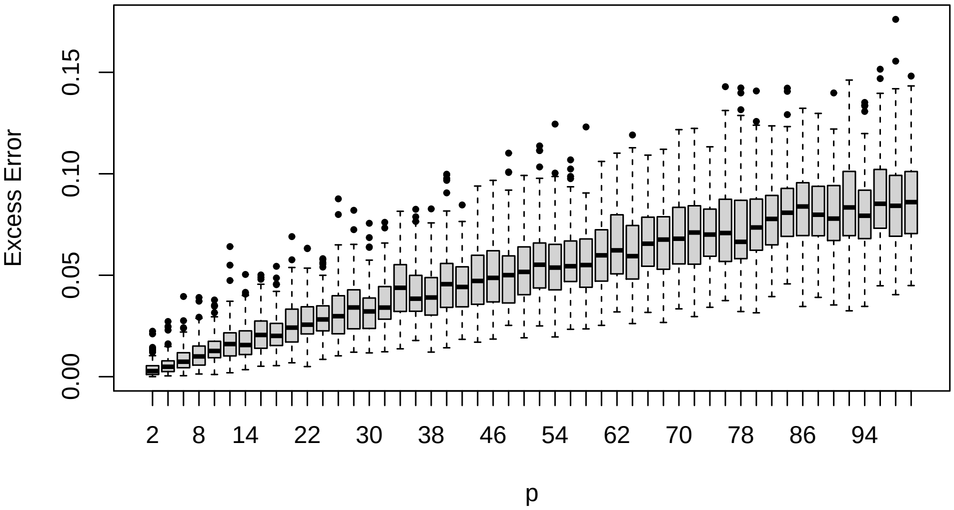

Figure 5: Boxplots of the excess error rates of diagonal LDA from 100 replicates of simulated data of size $n = 100$ generated from a balanced LDA setting with $\Delta = 2$ as the input dimension varies from 2 to 100.

6.1 Dense Scenario with Independent Predictors

To examine the effect of the input dimension p on error rates of the diagonal LDA (DLDA) with a fixed sample size, we varied p from 2 to 100 and set the mean vectors for two classes to $\boldsymbol{\mu}_+ = \frac{\Delta}{2\sqrt{p}}1_p$ and $\boldsymbol{\mu}_- = -\frac{\Delta}{2\sqrt{p}}1_p$, respectively and the variance matrix to the identity matrix in a normal distribution setting with equal proportions, where 1_p is the p-vector of ones. Note that this setting gives the Mahalanobis distance of Δ for each p. 100 replicates of a sample of size 100 were generated from the specified normal distribution with $\Delta = 2$. To each replicate, we applied DLDA and calculated the error rate of the estimated rule analytically. Figure 5 shows boxplots of the excess error rates of DLDA as the input dimension grows from 2 to 100. The excess error rates appear to increase linearly with p in this setting. The size of the error rates indicates that it is feasible to discriminate the two classes with the simple procedure when the data dimensions are as high as the sample size.

For comparison, we applied SVM and penalized logistic regression with ℓ_2 norm to the samples with $p = 100$. For SVM, the R package `svmpath` was used and for penalized logistic regression, `glmnet` was used. Figure 6 shows the excess error curves calculated analytically for 100 replicates of estimated prediction rules as a function of penalty parameter λ. The left panel is for penalized logistic regression and the right panel is for SVM. The light lines are sample-specific error curves and the thick lines indicate the average errors. The error curves exhibit a typical pattern of the performance of penalized methods producing underfit to overfit rules as the penalty diminishes. The minimum excess error from the mean curves is around 0.1 for both methods, which suggests that they perform similarly if λ is tuned properly. The first two rows of Table 4 under the heading 'dense scenario' gives a numerical summary of the excess errors of the two methods based on the 100 samples when the penalty parameter is optimally chosen to minimize the prediction error for each sample. In terms

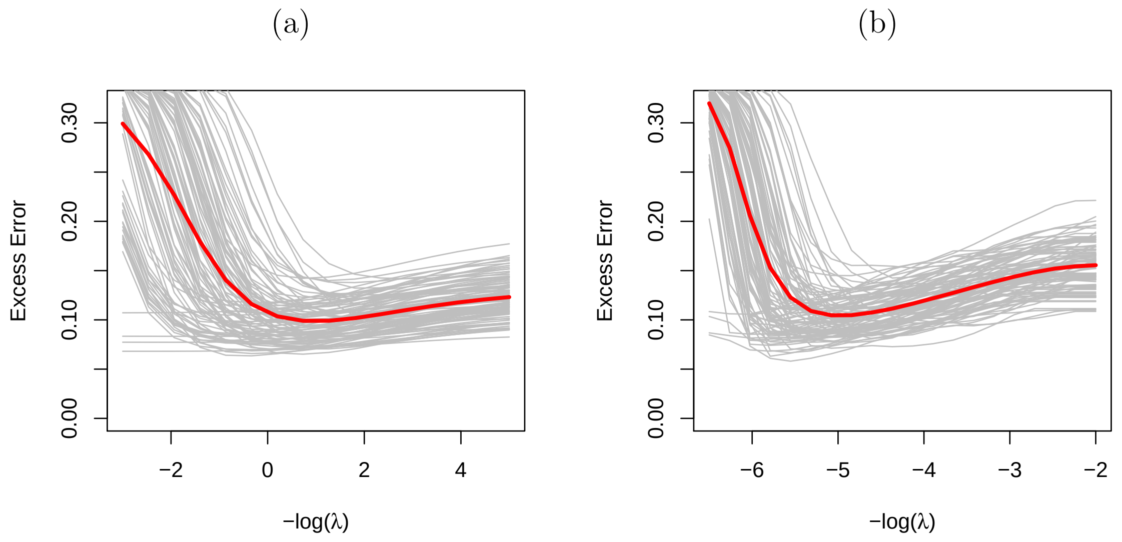

Figure 6: Excess error rate curves for estimated discriminant rules from (a) penalized logistic regression with ℓ_2 norm; (b) support vector machine when samples of size 100 are simulated from 100 dimensional LDA setting with equal class proportions and $\Delta = 2$. The light curves are sample-specific and the thick lines are the average over 100 samples.

of the mean error, logistic regression is slightly better than SVM. The table also provides the mean excess errors of DLDA and weighted LDA (WLDA), which is another version of regularized LDA with a convex combination of the estimated diagonal covariance matrix and the sample covariance matrix. Comparisons with DLDA and WLDA show that in this setting, the regularized versions of LDA give lower error rates than the two methods on average. However, it is no longer clear whether this observed advantage of LDA is due to difference in the mode of regularization or additional modeling.

6.2 Sparse Scenario with Varying Degree of Correlation

The canonical setting posits a very different scenario from the situation in the previous section where every dimension adds information about class discrimination. There may be a differential impact on the error rates of the methods for large p if $\boldsymbol{\mu}_+ = \frac{\Delta}{2}\mathbf{e}_1$ and $\boldsymbol{\mu}_- = -\frac{\Delta}{2}\mathbf{e}_1$, and thus additional predictors only increase dimensionality without adding information for classification. In addition to the potential impact of sparsity, we consider scenarios with correlated predictors by noting that independence of predictors would favor DLDA. A simple structure is assumed for covariance $\Sigma = [\sigma_{ij}]$, where $\sigma_{ii} = 1$ and $\sigma_{ij} = \rho$ for $i \neq j$. To keep the intrinsic level of difficulty the same across different settings, we set the mean vectors $\boldsymbol{\mu}_+ = \frac{\Delta}{2\sqrt{\mathbf{e}_1'\Sigma^{-1}\mathbf{e}_1}}\mathbf{e}_1$ and $\boldsymbol{\mu}_- = -\frac{\Delta}{2\sqrt{\mathbf{e}_1'\Sigma^{-1}\mathbf{e}_1}}\mathbf{e}_1$ in general so that the Mahalanobis distance is Δ.

The lower block of Table 4 under the heading ‘sparse scenario’ gives comparisons of the methods for this sparse LDA setting with varying degree of correlation among the predictors. Again, $\Delta = 2$, $p = 100$ and the summary statistics are based on 100 replicates. The performance of penalized logistic regression appears markedly different between the dense and sparse scenarios. The rest of the methods do not seem to be affected by sparsity in the mean vectors. Clearly as the extent of correlation increases, DLDA becomes ineffective. The lowest errors are achieved when the convex combination of covariance matrices in WLDA is optimized for the best prediction. In contrast to logistic regression, SVM shows stable performance across a wide range of settings.

Table 4: Mean and (standard deviation) of the excess error rates of diagonal LDA, weighted LDA, penalized logistic regression and SVM with optimally chosen penalty parameters for samples of size 100 from high dimensional LDA settings with $p = 100$

Scenario	DLDA	WLDA	Logistic Regression	SVM
Dense	0.0872	0.0864	0.0949	0.0979
	(0.0133)	(0.0134)	(0.0186)	(0.0187)
Sparse				
$\rho = 0$	0.0854	0.0850	0.1572	0.0968
	(0.0200)	(0.0200)	(0.0165)	(0.0186)
$\rho = 0.1$	0.1214	0.0929	0.1549	0.1013
	(0.0417)	(0.0215)	(0.0170)	(0.0177)
$\rho = 0.3$	0.1978	0.0951	0.1453	0.1028
	(0.0711)	(0.0213)	(0.0174)	(0.0175)
$\rho = 0.5$	0.2452	0.0958	0.1341	0.1030
	(0.0772)	(0.0207)	(0.0177)	(0.0175)

7 Conclusion

This paper has shown that many popular classification methods can be compared analytically in terms of the efficiency in reducing error rates, using standard asymptotic techniques. Though the results are obtained under a special setting where clean analysis is feasible, they lead to interesting theoretical comparisons of the methods and shed light on their relative merits and drawbacks. When modeling approach is compared with algorithmic approach under the normal setting, it is found that modeling generally leads to more efficient use of data. In particular, the SVM is shown to be between 40% and 67% as effective as LDA while boosting is between 20% and 54% as effective as LDA, when the Bayes error rate ranges from 4% to 10%. However, a loss function plays an important role in determining the efficiency of the corresponding procedure. The smooth SVM with squared hinge loss turns out to be more effective than logistic regression under the normal setting.

Since the correct form of a model is not known a priori in practice, it is important to understand the impact of model misspecification on the error rate. The comparisons under the QDA setting and the LDA setting with label noise indicate that there is a trade-off between efficiency and robustness.

The theoretical comparisons presented in this paper can be extended in many directions. To extend the scope of comparison to more complex settings with a nonlinear boundary, various probability models for two classes can be considered together with expanding families for discriminant functions. Possible models include the QDA setting with quadratic discriminant functions as an immediate extension and a mixture of several Gaussian components for each class. Further nonparametric generalization can be achieved via expansion of discriminant features, either by basis expansion or feature mapping through a kernel. Undoubtedly, analytical comparisons will become increasingly complex for nonparametric setting.

Another direction of extension is to allow the dimension p to grow with sample size n. In the current analysis, the dimension of the attributes p is assumed fixed, and the limiting behavior of a discriminant function is examined as n goes to ∞. In the classical asymptotics setting, a probability model of modest complexity can be estimated reasonably well with a sufficient number of observations. However, one of the main challenges faced in modern data analysis is high dimensionality of data. Practical successes of such prediction-oriented classification procedures as the SVM and boosting partly lie in their ability to handle high dimensional features. In one of the earlier references of the SVM, Cortes and Vapnik (1995) noted how quickly the number of parameters to estimate increases in Fisher's normal discriminant paradigm as the dimension of the feature space increases, and proposed to aim at classification boundary directly, instead of probability model parameters. To cope with the dimensionality, it is necessary to cast the procedures in a regularization framework for both technical and computational reasons. It would be interesting to extend the current analysis to high dimensional setting with proper regularization of discriminant coefficients and to study the effect of a different style of regularization on the error rate and efficiency. Theory of penalized M-estimators will be useful for comparisons of relative merits of the competing procedures; see, for example, Bühlmann and Van De Geer (2011), and see also Bickel and Levina (2004) and Rocha et al. (2009) for related results.

Appendix

This appendix contains derivations of the optimal discriminant coefficients and the asymptotic distribution of a coefficient vector for variants of the SVM and boosting in Section 3.

Smooth SVM

The risk of $\underline{\boldsymbol{\beta}}$ under the squared hinge loss in the LDA setting is given by

$$L(\underline{\boldsymbol{\beta}}) = \pi_+\sigma^2\left[(a_p^2+1)\Phi(a_p)+a_p\phi(a_p)\right] + \pi_-\sigma^2\left[(a_m^2+1)\Phi(a_m)+a_m\phi(a_m)\right],$$

where $\sigma \equiv \sqrt{\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}}$, and a_p and a_m are defined in (9).

To identify the optimal coefficients $\underline{\boldsymbol{\beta}}^*$, we take the derivatives of $L(\underline{\boldsymbol{\beta}})$ and equate them to zero:

$$\begin{aligned}
\frac{\partial L(\underline{\boldsymbol{\beta}})}{\partial \beta_0} &= -2\pi_+\sigma\left[a_p\Phi(a_p)+\phi(a_p)\right] + 2\pi_-\sigma\left[a_m\Phi(a_m)+\phi(a_m)\right] = 0 \\
\frac{\partial L(\underline{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} &= -2\pi_+\sigma\left[a_p\Phi(a_p)+\phi(a_p)\right]\boldsymbol{\mu}_+ + 2\pi_+\Phi(a_p)\boldsymbol{\Sigma}\boldsymbol{\beta} \\
&\quad +2\pi_-\sigma\left[a_m\Phi(a_m)+\phi(a_m)\right]\boldsymbol{\mu}_- + 2\pi_-\Phi(a_m)\boldsymbol{\Sigma}\boldsymbol{\beta} = \mathbf{0}.
\end{aligned}$$

Letting $\Theta(z) \equiv z\Phi(z)+\phi(z)$ and $\omega^* \equiv \frac{\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}^*}{\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}^*\|}$, we can simplify the above equations to

$$\begin{aligned}
\pi_+\Theta(a_p) &= \pi_-\Theta(a_m) \qquad (22) \\
\left[\pi_+\Phi(a_p)+\pi_-\Phi(a_m)\right]\boldsymbol{\Sigma}^{\frac{1}{2}}\omega^* &= \pi_+\Theta(a_p)\boldsymbol{\mu}_+ - \pi_-\Theta(a_m)\boldsymbol{\mu}_-. \qquad (23)
\end{aligned}$$

By solving the equations for $\underline{\boldsymbol{\beta}}^*$, we have the optimal parameters

$$\beta_0^* = \frac{a_m - a_p}{a_p + a_m + \Delta} - \frac{1}{2}(\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-)'\boldsymbol{\beta}^* \text{ and } \boldsymbol{\beta}^* = \frac{2}{[a_p + a_m + \Delta]\Delta}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-),$$

where a_p and a_m are the constants that solve the following equations:

$$\begin{aligned}
\pi_+\Theta(a_p) &= \pi_-\Theta(a_m) \\
\pi_+\Theta(a_p)\Delta &= \pi_+\Phi(a_p) + \pi_-\Phi(a_m).
\end{aligned}$$

The second equation comes from the fact that ω^* in (23) is a unit vector.

In the canonical LDA setting with equal probabilities, we can verify that

$$H(\underline{\boldsymbol{\beta}}^*) = \begin{pmatrix} 2\Phi(a^*) & \mathbf{0} \\ \mathbf{0} & 2\Phi(a^*)\mathbf{I}_p + M\mathbf{J}_p \end{pmatrix},$$

where $\mathbf{J}_p = \mathbf{e}_1\mathbf{e}_1'$ and $M = \frac{\Delta^2}{2}\Phi(a^*) - 2(a^* + \Delta)\phi(a^*)$, and

$$G(\underline{\boldsymbol{\beta}}^*) = 8\ c_{\text{SSVM}}\Phi(a^*)\begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_p + \frac{\Delta^2}{4}\mathbf{J}_p - c_{\text{SSVM}}\Delta^2\mathbf{J}_p \end{pmatrix}.$$

Hence the limiting covariance matrix is

$$H(\boldsymbol{\beta}^*)^{-1}G(\boldsymbol{\beta}^*)H(\boldsymbol{\beta}^*)^{-1} = c_{\text{SSVM}}^2\begin{pmatrix} \kappa_1 & 0 & \cdots & & 0 \\ 0 & \kappa_2 & & & \\ \vdots & & \kappa_1 & & \\ & & & \ddots & \\ 0 & & & & \kappa_1 \end{pmatrix}$$

where $\kappa_1 = \dfrac{(2a^* + \Delta)\Delta}{\Phi(a^*)} = \dfrac{2}{c_{\text{SSVM}}\Phi(a^*)}$ and $\kappa_2 = \dfrac{8\left(1 + \frac{1}{4}\Delta^2 - c_{\text{SSVM}}\Delta^2\right)\Phi(a^*)}{c_{\text{SSVM}}\left[2\Phi(a^*) + M\right]^2}$.

Huberized SVM

With some additional definition of constants depending on k, we have the true risk of β_0 and $\boldsymbol{\beta}$ under the ρ_k in the LDA setting expressed as

$$\begin{aligned}
L(\underline{\boldsymbol{\beta}}) \;\; = \;\; & \pi_+\sigma^2\left[(a_p^2+1)\Phi(a_p)+a_p\phi(a_p)\right]+\pi_-\sigma^2\left[(a_m^2+1)\Phi(a_m)+a_m\phi(a_m)\right] \\
& +\pi_+\Phi(a_p^k)\left[(1-k)(a_p+a_p^k)\sigma-(a_p^2+1)\sigma^2\right]+\pi_-\Phi(a_m^k)\left[(1-k)(a_m+a_m^k)\sigma-(a_m^2+1)\sigma^2\right] \\
& -\pi_+\phi(a_p^k)a_p^k\sigma^2-\pi_-\phi(a_m^k)a_m^k\sigma^2,
\end{aligned}$$

where $a_p^k \equiv \dfrac{k-\beta_0-\boldsymbol{\beta}'\boldsymbol{\mu}_+}{\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}\|}$, and $a_m^k \equiv \dfrac{k+\beta_0+\boldsymbol{\beta}'\boldsymbol{\mu}_-}{\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}\|}$. Then

$$\begin{aligned}
\frac{\partial L(\underline{\boldsymbol{\beta}})}{\partial \beta_0} = & -2\pi_+\sigma\left[a_p\Phi(a_p)+\phi(a_p)-a_p^k\Phi(a_p^k)-\phi(a_p^k)\right] \\
& +2\pi_-\sigma\left[a_m\Phi(a_m)+\phi(a_m)-a_m^k\Phi(a_m^k)-\phi(a_m^k)\right], \text{ and}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial L(\underline{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} = & -2\pi_+\sigma\left[a_p\Phi(a_p)+\phi(a_p)-a_p^k\Phi(a_p^k)-\phi(a_p^k)\right]\boldsymbol{\mu}_+ + 2\pi_+\left[\Phi(a_p)-\Phi(a_p^k)\right]\boldsymbol{\Sigma}\boldsymbol{\beta} \\
& +2\pi_-\sigma\left[a_m\Phi(a_m)+\phi(a_m)-a_m^k\Phi(a_m^k)-\phi(a_m^k)\right]\boldsymbol{\mu}_- + 2\pi_-\left[\Phi(a_m)-\Phi(a_m^k)\right]\boldsymbol{\Sigma}\boldsymbol{\beta}.
\end{aligned}$$

With the earlier definition of $\Theta(z)=z\Phi(z)+\phi(z)$ and $\omega^* \equiv \frac{\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}^*}{\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}^*\|}$, we can show that the optimality equation $S(\underline{\boldsymbol{\beta}})=\mathbf{0}$ becomes

$$\begin{aligned}
\pi_+[\Theta(a_p)-\Theta(a_p^k)] &= \pi_-[\Theta(a_m)-\Theta(a_m^k)] \\
\left[\pi_+\{\Phi(a_p)-\Phi(a_p^k)\}+\pi_-\{\Phi(a_m)-\Phi(a_m^k)\}\right]\boldsymbol{\Sigma}^{\frac{1}{2}}\omega^* &= \pi_+[\Theta(a_p)-\Theta(a_p^k)]\boldsymbol{\mu}_+ - \pi_-[\Theta(a_m)-\Theta(a_m^k)]\boldsymbol{\mu}_-.
\end{aligned}$$

Similar to the smooth SVM, the optimal parameters are expressed as

$$\beta_0^* = \frac{a_m-a_p}{a_p+a_m+\Delta}-\frac{1}{2}(\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-)'\boldsymbol{\beta}^* \text{ and } \boldsymbol{\beta}^* = \frac{2}{[a_p+a_m+\Delta]\Delta}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-),$$

where a_p and a_m are now defined as the constants that solve the following equations:

$$\begin{aligned}
\pi_+[\Theta(a_p)-\Theta(a_p^k)] &= \pi_-[\Theta(a_m)-\Theta(a_m^k)] \\
\pi_+[\Theta(a_p)-\Theta(a_p^k)]\Delta &= \pi_+[\Phi(a_p)-\Phi(a_p^k)]+\pi_-[\Phi(a_m)-\Phi(a_m^k)].
\end{aligned}$$

Note the relations that

$$a_p^k = a_p - \frac{1-k}{\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}^*\|}, \quad a_m^k = a_m - \frac{1-k}{\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}^*\|}, \quad \text{and} \quad \|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}^*\| = \frac{2}{a_p+a_m+\Delta}.$$

Under the balanced canonical LDA setting, the Hessian matrix H is given by

$$H(\underline{\boldsymbol{\beta}}^*) = \begin{pmatrix} 2[\Phi(a^*)-\Phi(a_k^*)] & 0 \\ 0 & 2\left[\Phi(a^*)-\Phi(a_k^*)\right]\mathbf{I}_p + M_k\mathbf{J}_p \end{pmatrix},$$

where $M_k = \dfrac{\Delta^2}{2}\left[\Phi(a^*) - \Phi(a_k^*)\right] - 2\Delta\left[\phi(a^*) - \phi(a_k^*)\right] - 2\left[a^*\phi(a^*) - a_k^*\phi(a_k^*)\right]$. Having the hybrid of a quadratic loss in the interval from k to 1 and a linear loss below k brings a slight change in the form of the H as compared to the H matrix of the smooth SVM. With similar changes in the elements, G matrix is given by

$$G(\boldsymbol{\beta}^*) = 8c_{\text{HSVM}}[\Phi(a^*) - \Phi(a_k^*) + \frac{\Delta}{2}(k-1)\Theta(a_k^*)] \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{I}_p + \frac{1}{4}\Delta^2\mathbf{J}_p \end{pmatrix} - 8c_{\text{HSVM}}D_k \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{J}_p \end{pmatrix},$$

where

$$D_k = c_{\text{HSVM}}\Delta^2\left[\Delta\{a^*(\Phi(a^*) - \Phi(a_k^*)) - (\phi(a^*) - \phi(a_k^*))\} - (\Phi(a^*) - \Phi(a_k^*))\right] - \Delta(k-1)\phi(a_k^*).$$

Note that when k goes to $-\infty$, $H(\boldsymbol{\beta}^*)$ and $G(\boldsymbol{\beta}^*)$ for the Huberized SVM above reduce to those for the smooth SVM. The limiting covariance matrix is

$$H(\boldsymbol{\beta}^*)^{-1}G(\boldsymbol{\beta}^*)H(\boldsymbol{\beta}^*)^{-1} = c_{\text{HSVM}}^2 \begin{pmatrix} \kappa_1 & 0 & & \cdots & & 0 \\ 0 & \kappa_2 & & & & \\ & & \kappa_1 & \iddots & & \vdots \\ \vdots & & \ddots & \ddots & & \\ & & & & \ddots & 0 \\ 0 & & \cdots & & 0 & \kappa_1 \end{pmatrix},$$

where

$$\begin{aligned} \kappa_1 &= \frac{2[\Phi(a^*) - \Phi(a_k^*) + \frac{\Delta}{2}(k-1)\Theta(a_k^*)]}{c_{\text{HSVM}}\left[\Phi(a^*) - \Phi(a_k^*)\right]^2} \quad \text{and} \\ \kappa_2 &= \frac{8\Big((1 + \frac{1}{4}\Delta^2)\{\Phi(a^*) - \Phi(a_k^*) + \frac{\Delta}{2}(k-1)\Theta(a_k^*)\} - D_k\Big)}{c_{\text{HSVM}}\left[2(\Phi(a^*) - \Phi(a_k^*)) + M_k\right]^2}. \end{aligned}$$

Boosting

Under the LDA setting in (1), the true risk is given by

$$L(\underline{\boldsymbol{\beta}}) = \pi_+ \exp\big(-\beta_0 - \boldsymbol{\beta}'\boldsymbol{\mu}_+ + \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}\big) + \pi_- \exp\big(\beta_0 + \boldsymbol{\beta}'\boldsymbol{\mu}_- + \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}\big).$$

Equating the gradient of $L(\underline{\boldsymbol{\beta}})$ to $\mathbf{0}$ for the population minimizer $\underline{\boldsymbol{\beta}}^*$, we have

$$\begin{aligned} S(\underline{\boldsymbol{\beta}}) = \pi_+ \exp(-\beta_0 - \boldsymbol{\beta}'\boldsymbol{\mu}_+ + \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}) &\begin{pmatrix} -1 \\ -\boldsymbol{\mu}_+ + \boldsymbol{\Sigma}\boldsymbol{\beta} \end{pmatrix} \\ + \pi_- \exp(\beta_0 + \boldsymbol{\beta}'\boldsymbol{\mu}_- + \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}) &\begin{pmatrix} 1 \\ \boldsymbol{\mu}_- + \boldsymbol{\Sigma}\boldsymbol{\beta} \end{pmatrix} = \mathbf{0}. \end{aligned}$$

By solving the equation for $\underline{\boldsymbol{\beta}}$, we have

$$\beta_0^* = \frac{1}{2}\left(\log\frac{\pi_+}{\pi_-} - \frac{1}{2}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-)\right) \text{ and } \boldsymbol{\beta}^* = \frac{1}{2}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-).$$

The H and G matrices under the exponential loss in the general LDA setting are given by

$$H(\underline{\boldsymbol{\beta}}^*) = \sqrt{\pi_+\pi_-}\exp\left\{-\frac{1}{8}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)\right\}\begin{pmatrix} 2 & (\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-)' \\ \boldsymbol{\mu}_+ + \boldsymbol{\mu}_- & 2\boldsymbol{\Sigma} + (\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-)(\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-)' \end{pmatrix}$$

and

$$G(\underline{\boldsymbol{\beta}}^*) = \begin{pmatrix} 1 & (\pi_+\boldsymbol{\mu}_+ + \pi_-\boldsymbol{\mu}_-)' \\ \pi_+\boldsymbol{\mu}_+ + \pi_-\boldsymbol{\mu}_- & \boldsymbol{\Sigma} + \pi_+\boldsymbol{\mu}_+\boldsymbol{\mu}_+' + \pi_-\boldsymbol{\mu}_-\boldsymbol{\mu}_-' \end{pmatrix}.$$

Under the canonical LDA setting (2), they are simplified to

$$H(\underline{\boldsymbol{\beta}}^*) = 2\sqrt{\pi_+\pi_-}\exp\left(-\frac{\Delta^2}{8}\right)\mathbf{I}_{p+1}, \text{ and } G(\underline{\boldsymbol{\beta}}^*) = \begin{pmatrix} 1 & \frac{\Delta}{2}(\pi_+ - \pi_-)\mathbf{e}_1' \\ \frac{\Delta}{2}(\pi_+ - \pi_-)\mathbf{e}_1 & \mathbf{I}_p + \frac{\Delta^2}{4}\mathbf{J}_p \end{pmatrix},$$

where $\mathbf{J}_p = \mathbf{e}_1\mathbf{e}_1'$. Further, equal class proportions yield

$$H(\underline{\boldsymbol{\beta}}^*) = \exp\left(-\frac{\Delta^2}{8}\right)\mathbf{I}_{p+1}, \text{ and } G(\underline{\boldsymbol{\beta}}^*) = \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{I}_p + \frac{\Delta^2}{4}\mathbf{J}_p \end{pmatrix}.$$

Hence the limiting covariance matrix of the discriminant coefficient vector is

$$H(\underline{\boldsymbol{\beta}}^*)^{-1}G(\underline{\boldsymbol{\beta}}^*)H(\underline{\boldsymbol{\beta}}^*)^{-1} = \exp\left(\frac{\Delta^2}{4}\right)\begin{pmatrix} 1 & 0 & \cdots & & 0 \\ 0 & 1+\frac{\Delta^2}{4} & & & \\ \vdots & & 1 & & \\ & & & \ddots & \\ 0 & & & & 1 \end{pmatrix}.$$

References

Bartlett, P., Jordan, M. and McAuliffe, J. (2006). Convexity, classification, and risk bounds, *Journal of the American Statistical Association* **101**: 138–156.

Bickel, P. J. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations, *Bernoulli* **10**(6): 989–1010.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Series in Statistics, Springer.

Cortes, C. and Vapnik, V. (1995). Support-Vector Networks, *Machine Learning* **20**(3): 273–297.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*, Cambridge University Press.

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*, Springer.

Duda, R. O., Hart, P. E. and Stork, D. G. (2000). *Pattern Classification (2nd Edition)*, Wiley-Interscience.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis, *Journal of the American Statistical Association* **70**(352): 892–898.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences* **55**(1): 119–139.

Friedman, J., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting, *The Annals of Statististics* **28**: 337–407.

Geyer, C. J. (1994). On the asymptotics of constrained M-estimation, *The Annals of Statistics* **22**: 1993–2010.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*, Springer Verlag, New York.

Hjort, N. L. and Pollard, D. (1993). Asymptotics for minimisers of convex processes, *Technical report*, Department of Statistics, Yale University.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators, *The Annals of Statistics* **28**(5): 1356–1378.

Koo, J.-Y., Lee, Y., Kim, Y. and Park, C. (2008). A Bahadur representation of the linear Support Vector Machine, *Journal of Machine Learning Research* **9**: 1343–1368.

Lee, Y.-J. and Mangasarian, O. (2001). SSVM: A smooth support vector machine, *Computational Optimization and Applications* **20**: 5–22.

Lin, Y. (2002). A note on margin-based loss functions in classification, *Statististics and Probability Letters* **68**: 73–82.

McLachlan, G. J. (2004). *Discriminant Analysis and Statistical Pattern Recognition*, Wiley-Interscience.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators, *Econometric Theory* **7**: 186–199.

Rocha, G., Wang, X. and Yu, B. (2009). Asymptotic distribution and sparsistency for l_1 penalized parametric M-estimators, with applications to linear SVM and logistic regression, *arXiv* **0908.1940v1**: 1–55.

Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths, *The Annals of Statistics* **35**(3): 1012–1030.

Schölkopf, B. and Smola, A. (2002). *Learning with Kernels - Support Vector Machines, Regularization, Optimization and Beyond*, Cambridge, MA: MIT Press.

Steinwart, I. (2005). Consistency of support vector machines and other regularized kernel machines, *IEEE Transactions on Information Theory* **51**: 128–142.

van der Vaart, A. W. (2000). *Asymptotic statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Vapnik, V. (1998). *Statistical Learning Theory*, Wiley, New York.

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization, *Annals of Statistics* **32**(1): 56–85.