Comment on "On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes"

Jing-Hao Xue (jinghao@stats.gla.ac.uk) and D. Michael Titterington (mike@stats.gla.ac.uk)

Department of Statistics, University of Glasgow, Glasgow G12 8QQ, UK

Abstract. Comparison of generative and discriminative classifiers is an ever-lasting topic. As an important contribution to this topic, based on their theoretical and empirical comparisons between the naïve Bayes classifier and linear logistic regression, Ref. [6] claimed that there exist two distinct regimes of performance between the generative and discriminative classifiers with regard to the training-set size. In this paper, our empirical and simulation studies, as a complement of their work, however, suggest that the existence of the two distinct regimes may not be so reliable. In addition, for real world datasets, so far there is no theoretically correct, general criterion for choosing between the discriminative and the generative approaches to classification of an observation $\mathbf{x}$ into a class y; the choice depends on the relative confidence we have in the correctness of the specification of either $p(y|\mathbf{x})$ or $p(\mathbf{x}, y)$ for the data. This can be to some extent a demonstration of why Ref. [3] and [7] prefer normal-based linear discriminant analysis (LDA) when no model mis-specification occurs but other empirical studies may prefer linear logistic regression instead. Furthermore, we suggest that pairing of either LDA assuming a common diagonal covariance matrix (LDA-Λ) or the naïve Bayes classifier and linear logistic regression may not be perfect, and hence it may not be reliable for any claim that was derived from the comparison between LDA-Λ or the naïve Bayes classifier and linear logistic regression to be generalised to all generative and discriminative classifiers.

Keywords: Asymptotic relative efficiency; Discriminative classifiers; Generative classifiers; Logistic regression; Normal-based discriminant analysis; Naïve Bayes classifier

Abbreviations: LDA/QDA – Normal-based Linear/Quadratic Discriminant Analysis; AIC – Akaike Information Criterion; GAM – Generalised Additive Model

1. Introduction

Classification is a ubiquitous problem tackled in statistics, machine learning, pattern recognition and data mining [4].

Generative classifiers, also termed the sampling paradigm [2], such as normal-based discriminant analysis and the naïve Bayes classifier, model the joint distribution $p(\mathbf{x}, y)$ of the measured features $\mathbf{x}$ and the class labels y factorised in the form $p(\mathbf{x}|y)p(y)$, and learn the model parameters through maximisation of the likelihood given by $p(\mathbf{x}|y)p(y)$.

Discriminative classifiers, also termed the diagnostic paradigm [2], such as logistic regression, model the conditional distribution $p(y|\mathbf{x})$ of the class labels given the features, and learn the model parameters through maximising the conditional likelihood based on $p(y|\mathbf{x})$.

Comparison of generative and discriminative classifiers is an everlasting topic [3,6,7,10,12]. Results from such comparisons, in particular in terms of misclassification error rates, can not only guide the selection of an appropriate classifier, either generative or discriminative, but also shed light on how to exploit the best of both worlds of classifiers, and thus has been attracting long-standing interest from both researchers and practitioners.

An important contribution to this topic is from Ref. [6], presenting some theoretical and empirical comparisons between linear logistic regression and the naïve Bayes classifier.

The results in [6] suggested that, between the two classifiers, there were two distinct regimes of discriminant performance with respect to the training-set size. More precisely, they proposed that the discriminative classifier had lower asymptotic error rate while the generative classifier may approach its (higher) asymptotic error rate much faster. In other words, the discriminative classifier performs better with larger training sets while the generative classifier does better with smaller training sets.

However, Ref. [3] and [7] presented some theoretical and simulation studies showing that normal-based linear discriminant analysis (LDA), a generative classifier, has better asymptotic efficiency (i.e., performs better with larger training sets) when no model mis-specification occurs. Our empirical and simulation studies, as presented in this paper, suggest that it may not be so reliable to claim such an existence of the two distinct regimes. Furthermore, we suggest that pairing of either LDA assuming a common diagonal covariance matrix Λ (denoted by LDA-Λ hereafter) or the naïve Bayes classifier and linear logistic regression may not be perfect, and hence it may not be reliable for any claim that was derived from the comparison between LDA-Λ or the naïve Bayes classifier and linear logistic regression to be generalised to all generative and discriminative classifiers.

2. Setting for Comparison

2.1. Setting used by Ref. [6]

The setting for the theoretical proof and empirical evidence in [6] includes a binary class label y, *e.g.*, $y \in \{1, 2\}$, a p-dimensional feature vector $\mathbf{x}$ and the assumption of conditional independence amongst $\mathbf{x}|y$, the features within a class.

The naïve Bayes classifier, a generative classifier defined as in equation (4) in Section 2.2, assumes statistically independent features $\mathbf{x}$ within classes y and thus diagonal covariance matrices within classes. By contrast, linear logistic regression, a discriminative classifier defined as in equation (1), may not assume such conditional independence of the components of $\mathbf{x}$. Both classifiers can be applied to discrete, continuous or mixed-valued features $\mathbf{x}$.

In the case of discrete features $\mathbf{x}$, each feature $x_i, i = 1, \ldots, p$, independently of other features within $\mathbf{x}$, is assumed within a class to be a binomial variable such that its value $x_i \in \{0, 1\}$. However, this may not guarantee the discriminant function $\lambda(\alpha) = \log\{p(y = 1|\mathbf{x})/p(y = 2|\mathbf{x})\}$ of the naïve Bayes classifier, where α is a parameter vector, to be linear. As linear logistic regression uses a linear discriminant function, the naïve Bayes classifier may not be a partner of linear logistic regression as a generative-discriminative pair (see Section 2.2 for more discussion about this pairing).

In the case of continuous features $\mathbf{x}$, $\mathbf{x}|y$ is assumed to follow Gaussian distributions with equal covariance matrices across the two classes, *i.e.*, $\Sigma_1 = \Sigma_2$ and, in view of the conditional independence assumption, both covariance matrices are equal to a diagonal matrix Λ. Some algebra shows that, under the assumption of a common diagonal covariance matrix Λ for normally distributed data, the naïve Bayes method is equivalent to LDA-Λ (defined as equation (2)), and, under the assumption of unequal diagonal within-class covariance matrices, it is equivalent to quadratic discriminant analysis. For the experiments in [6], all of the observed values of the features are rescaled so that $x_i \in [0, 1]$.

Based on such a setting, Ref. [6] compared two so-called generative-discriminative pairs: one is for the continuous case, comparing LDA assuming a common diagonal covariance matrix (LDA-Λ) vs. linear logistic regression, and the other is for the discrete case, comparing the naïve Bayes classifier vs. linear logistic regression. We shall next make some comments on these pairings.

2.2. On the Pairing of LDA-Λ/Naïve Bayes and Linear Logistic Regression/GAM

As mentioned in Section 2.1, first, the naïve Bayes classifier cannot guarantee the linear form of the discriminant function $\lambda(\alpha) = \log\{p(y = 1|\mathbf{x})/p(y = 2|\mathbf{x})\}$, and, secondly, the conditional independence amongst the multiple features within a class is a necessary condition for the validity of the naïve Bayes classifier and LDA-Λ but not for linear logistic regression, although in the latter the discriminant function $\lambda(\alpha)$ is modelled as a linear combination of separate features. Therefore, the comparison between a generative-discriminative pair of LDA-Λ/naïve Bayes classifier vs. linear logistic regression should be interpreted with caution, in particular when the data do not support the assumption of conditional independence of $\mathbf{x}|y$ that may shed unfavourable light on the simplified generative version, LDA-Λ and the naïve Bayes classifier.

In this section, we will illustrate two such generative-discriminative pairs: one is LDA-Λ vs. linear logistic regression [6], and the other is the naïve Bayes classifier vs. the generalised additive model (GAM) [10].

2.2.1. *LDA-Λ vs. Linear Logistic Regression*

Consider a feature vector $\mathbf{x} = (x_1, \ldots, x_p)^T$ and a binary class label $y = 1, 2$.

Linear logistic regression, one of the discriminative classifiers that do not assume any distribution $p(\mathbf{x}|y)$ of the data, is modelled directly with a linear discriminant function as

$$\lambda_{\text{dis}}(\alpha) = \log \frac{p(y=1|\mathbf{x})}{p(y=2|\mathbf{x})} = \log \frac{\pi_1}{\pi_2} + \log \frac{p(\mathbf{x}|y=1)}{p(\mathbf{x}|y=2)} = \beta_0 + \beta^T \mathbf{x} \ , \quad (1)$$

where $p(y = k) = \pi_k$, $\alpha^T = (\beta_0, \beta^T)$ and β is a parameter vector of p elements. By "linear", we mean a scalar-valued function of a linear combination of the features $x_1, \ldots, x_p$ of an observed feature vector $\mathbf{x}$.

By contrast, LDA-Λ, one of the generative classifiers, assumes that the data arise from two p-variate normal distributions with different means but the same diagonal covariance matrix such that $(\mathbf{x}|y = k; \theta) \sim \mathcal{N}(\mu_k, \Lambda)$, $k = 1, 2$, where $\theta = (\mu_k, \Lambda)$; this implies an assumption of conditional independence between any two features $x_i|y$ and $x_j|y$, $i \neq j$, within a class. The density function of $(\mathbf{x}|y = k; \theta)$ can be written as

$$p(\mathbf{x}|y=k;\theta) = \left\{ e^{\mu_k^T \Lambda^{-1} \mathbf{x}} \right\} \left\{ \frac{1}{\sqrt{(2\pi)^p |\Lambda|}} e^{-\frac{1}{2}\mu_k^T \Lambda^{-1} \mu_k} \right\} \left\{ e^{-\frac{1}{2}\mathbf{x}^T \Lambda^{-1} \mathbf{x}} \right\} \ ,$$

which leads to a linear discriminant function,

$$\lambda_{\text{gen}}(\alpha) = \log \frac{p(y=1|\mathbf{x})}{p(y=2|\mathbf{x})} = \log \frac{\pi_1}{\pi_2} + \log \frac{A(\theta_1, \eta)}{A(\theta_2, \eta)} + (\theta_1 - \theta_2)^T \mathbf{x} \ , \quad (2)$$

where $\theta_k = \mu_k^T \Lambda^{-1}$, $\eta = \Lambda^{-1}$ and $A(\theta_k, \eta) = \frac{1}{\sqrt{(2\pi)^p |\Lambda|}} e^{-\frac{1}{2}\mu_k^T \Lambda^{-1} \mu_k}$.

Similarly, by assuming that the data arise from two p-variate normal distributions with different means but the same full covariance matrix such that $(\mathbf{x}|y = k; \theta) \sim \mathcal{N}(\mu_k, \Sigma)$, $k = 1, 2$, we can obtain the same formula as $\lambda_{\text{gen}}(\alpha)$ but with $\theta_k = \mu_k^T \Sigma^{-1}$, $\eta = \Sigma^{-1}$ and $A(\theta_k, \eta) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} e^{-\frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k}$, which leads to the linear discriminant function of LDA with a common full covariance matrix Σ (denoted by LDA-Σ hereafter). Therefore, we could rewrite θ as $\theta = (\theta_k, \eta)$, where θ_k is a class-dependent parameter vector while η is a common parameter vector across the classes.

It is clear that the assumption of conditional independence amongst the features within a class is not a necessary condition for a generative classifier to attain a linear $\lambda_{\text{gen}}(\alpha)$. In fact, as pointed out by [7], if the feature vector $\mathbf{x}$ follows a multivariate exponential family distribution with the density or probability mass function within a class being

$$p(\mathbf{x}|y = k, \theta_k) = e^{\theta_k^T \mathbf{x}} A(\theta_k, \eta) h(\mathbf{x}, \eta), k = 1, 2 \ ,$$

the generative classifiers will attain a linear $\lambda_{\text{gen}}(\alpha)$.

2.2.2. *Naïve Bayes vs. Generalised Additive Model (GAM)*

As with logistic regression, a GAM does not assume any distribution $p(\mathbf{x}|y)$ for the data; it is modelled directly with a discriminant function that is a sum of p functions $f(x_i), i = 1, \ldots, p$, of the p features x_i

separately [10]; that is

$$\lambda_{\text{dis}}(\alpha) = \log \frac{p(y=1|\mathbf{x})}{p(y=2|\mathbf{x})} = \log \frac{\pi_1}{\pi_2} + \sum_{i=1}^{p} f(x_i) \ . \tag{3}$$

Meanwhile, along with the assumption of the distribution of $(\mathbf{x}|y)$, a fundamental assumption underlying the naïve Bayes classifier is that of the conditional independence amongst the p features within a class, so that the joint probability is $p(\mathbf{x}|y) = \prod_{i=1}^{p} p(x_i|y)$. It follows that the discriminant function $\lambda(\alpha)$ is

$$\lambda_{\text{gen}}(\alpha) = \log \frac{p(y=1|\mathbf{x})}{p(y=2|\mathbf{x})} = \log \frac{\pi_1}{\pi_2} + \sum_{i=1}^{p} \log \frac{p(x_i|y=1)}{p(x_i|y=2)} \ . \tag{4}$$

It is clear, as pointed out by [10], that the naïve Bayes classifier is a specialised case of a GAM, with $f(x_i) = \log\{p(x_i|y=1)/p(x_i|y=2)\}$. Furthermore, GAMs may not necessarily assume conditional independence.

One sufficient condition that leads to another specialised case of a GAM (we call it Q-GAM) is that $p(\mathbf{x}|y) = q(\mathbf{x}) \prod_{i=1}^{p} q(x_i|y)$, where $q(\mathbf{x})$ is common across the classes but cannot be further factorised into a product of functions of individual features as $\prod_{i=1}^{p} q(x_i)$. In such a case, the assumption of conditional independence between $x_i|y$ and $x_j|y$, $i \neq j$, is invalid but we still have $f(x_i) = \log\{q(x_i|y=1)/q(x_i|y=2)\}$, where $q(x_i|y)$ is different from the marginal probability $p(x_i|y)$ that is used by the naïve Bayes classifier.

2.2.3. *Summary*

First, the conditional independence amongst the features within a class is a necessary condition for the naïve Bayes classifier and LDA-Λ, but it is not a necessary condition for linear logistic regression. Therefore, the generative-discriminative pair of LDA with a common full covariance matrix Σ (LDA-Σ) vs. linear logistic regression also merits investigation.

Secondly, given the parity between $\lambda_{\text{gen}}(\alpha)$ and $\lambda_{\text{dis}}(\alpha)$ and thus that, between two pairs, LDA-Σ vs. linear logistic regression and Q-GAM vs. GAM in terms of classification, neither classifier assumes conditional independence of $\mathbf{x}|y$ amongst the features within a class, which is an elementary assumption underlying LDA-Λ and the naïve Bayes classifier. Therefore, it may not be reliable for any claim that is derived from the comparison between LDA-Λ or the naïve Bayes classifier and linear logistic regression to be generalised to all generative and discriminative classifiers.

Thirdly, a comparison of quadratic normal discriminant analysis (QDA) with unequal diagonal matrices Λ_1 and Λ_2 (denoted by QDA-Λ_g hereafter) and unequal full covariance matrices Σ_1 and Σ_2 (denoted by QDA-Σ_g hereafter) with quadratic logistic regression may provide an interesting extension of the work of [6].

2.3. Our implementation

Ref. [6] reported experimental results on 15 real-world datasets, 8 with only continuous and binary features and 7 with only discrete features, from the UCI machine learning repository [1]; this repository stores more than 100 datasets contributed and widely used by the machine learning community, as a benchmark for empirical studies of machine learning approaches. As pointed out in [6], there were a few cases (2 out of 8 continuous cases and 4 out of 7 discrete cases) that did not support the better asymptotic performance of the discriminative classifier, primarily because of the lack of sufficiently large training sets. However, it is known that the performance of a classifier varies to some extent with the features selected and a generally-valid empirical evaluation of classifiers is always an important but difficult problem [4]

In this context, we first replicate experiments on these 15 datasets, with and without stepwise variable selection being performed on the full linear logistic regression model using all the observations of each dataset. In the stepwise variable selection process, the decision to include or exclude a variable is based on the calculation of the Akaike information criterion (AIC). Furthermore, in the 8 continuous cases, both LDA-Λ and LDA-Σ are compared with linear logistic regression. Then we will extend the comparison to between QDA and quadratic logistic regression for the 8 continuous UCI datasets and finally to simulated continuous datasets.

The implementations in R (http://www.r-project.org/) of LDA and QDA are rewritten from a Matlab function *cda* for classical linear and quadratic discriminant analysis [13]. Logistic regression is implemented by an R function *glm* from a standard package **stats** in R, and the naïve Bayes classifier is implemented by an R function *naiveBayes* from a contributed package **e1071** for R.

In addition, similarly to what was done by [6], for each sampled training-set size m, we perform 1000 random splits of each dataset into a training set of size m and a test set of size $N - m$, where N is the number of observations in the whole dataset, and report the average of the misclassification error rates over these 1000 test sets. The training set is required to have at least 1 observation from each of the two classes, and, for discrete datasets, to have all the levels of the features presented by the training observations, otherwise the prediction for the test set may be asked to predict on some new levels for which no information has been provided in the training process.

In order to have all the coefficients of predictor variables in the model estimated in our implementation of logistic regression by *glm*, the number m of training observations should be larger than the number $\tilde{p}$ of predictor variables, where $\tilde{p} = p$ for the continuous cases if all p features are used for the linear model. More attention should be paid to the discrete cases with multinomial features in the model, where more dummy variables have to be used as the predictor variables, with the consequence that $\tilde{p}$ could be much larger than p, *e.g.*, $\tilde{p} = 3p$ for the linear model if all the features have 4 levels. In other words, although we may report misclassification error rates for logistic regression with small m, it is not reliable for us to base any general claim on those of m smaller than $\tilde{p}$, the actual number of predictor variables used by the logistic regression model.

3. Linear/Quadratic Discrimination for Empirical Datasets

3.1. Linear Discrimination for Continuous Datasets

For the continuous datasets, as was done by [6], all the multinomial features are removed so that only continuous and binary features x_i are kept and their values x_i are rescaled into $[0, 1]$. Any observation with missing features is removed from the datasets, as is any feature with only a single value for all the observations.

In addition, as Gaussian distributions and equal within-class covariance matrices are assumed for $\mathbf{x}|y$ for LDA-Λ and LDA-Σ, testing such assumptions can help the interpretation of classification performance of relevant classifiers. Therefore, before carrying out the classification, we perform the Shapiro-Wilk test for within-class normality for each feature $x_i|y$ [11] and Levene's test for homogeneity of variance across the two classes [5]. For the datasets discussed below, the significance level is set at 0.05, and we observe that null hypotheses of normality and homogeneity of variance are mostly rejected by the tests at that significance level.

A brief description of the continuous datasets can be found in Table I, which lists, for each dataset, the total number N_0 of the observations, the number N of the observations that we use after the pre-processing mentioned above, the total number p of continuous or binary features, the number p_{AIC} of features selected by AIC, the number p_{SW} of features for which the null hypotheses were rejected by the Shapiro-Wilk test and the corresponding number p_L for Levene's test, the indicator $\mathbf{1}_{\{2R-\Lambda\}} \in \{1, 0\}$ of whether or not the two regimes

Table I. Description of continuous datasets.

Dataset	N_0	N	p	p_{AIC}	p_{SW}	p_L	$\mathbf{1}_{\{2R-\Lambda\}}$	$\mathbf{1}_{\{2R-\Sigma\}}$
Pima	768	768	8	7	8	5	1	0
Adult	32561	1000	6	6	6	4	1	1
Boston	506	506	13	10	13	12	1	1
Optdigits 0-1	1125	1125	52	5	52	45	1	1
Optdigits 2-3	1129	1129	57	9	57	37	1	0
Ionosphere	351	351	33	20	33	27	1	1
Liver disorders	345	345	6	6	6	1	1	1
Sonar	208	208	60	37	59	16	1	1

are observed between LDA-Λ and linear logistic regression and the indicator $\mathbf{1}_{\{2R-\Sigma\}} \in \{1, 0\}$ with regard to LDA-Σ. Note that, for some large datasets such as “Adult” (and “Sick” in Section 3.3), in order to reduce computational complexity without degrading the validity of the comparison between the classifiers, we randomly sample observations with the class prior probability kept unchanged.

Our results are shown in Figure 1. Since with variable selection by AIC the results conform more to the claim of two regimes made by [6], we show such results only if they are different from those without variable selection. Meanwhile, in the figures hereafter we use the same annotation of the vertical and horizontal axes and the same line type as those in [6]. For the reason given at the end of Section 2.3, Figure 1 is only drawn for $m > p$, with the intercept in $\lambda(\alpha)$ taken into account.

In general, our study of these continuous datasets suggests the following conclusions.

First, in the comparison of LDA-Λ vs. linear logistic regression, the pattern of our results can be said to be similar to that of [6].

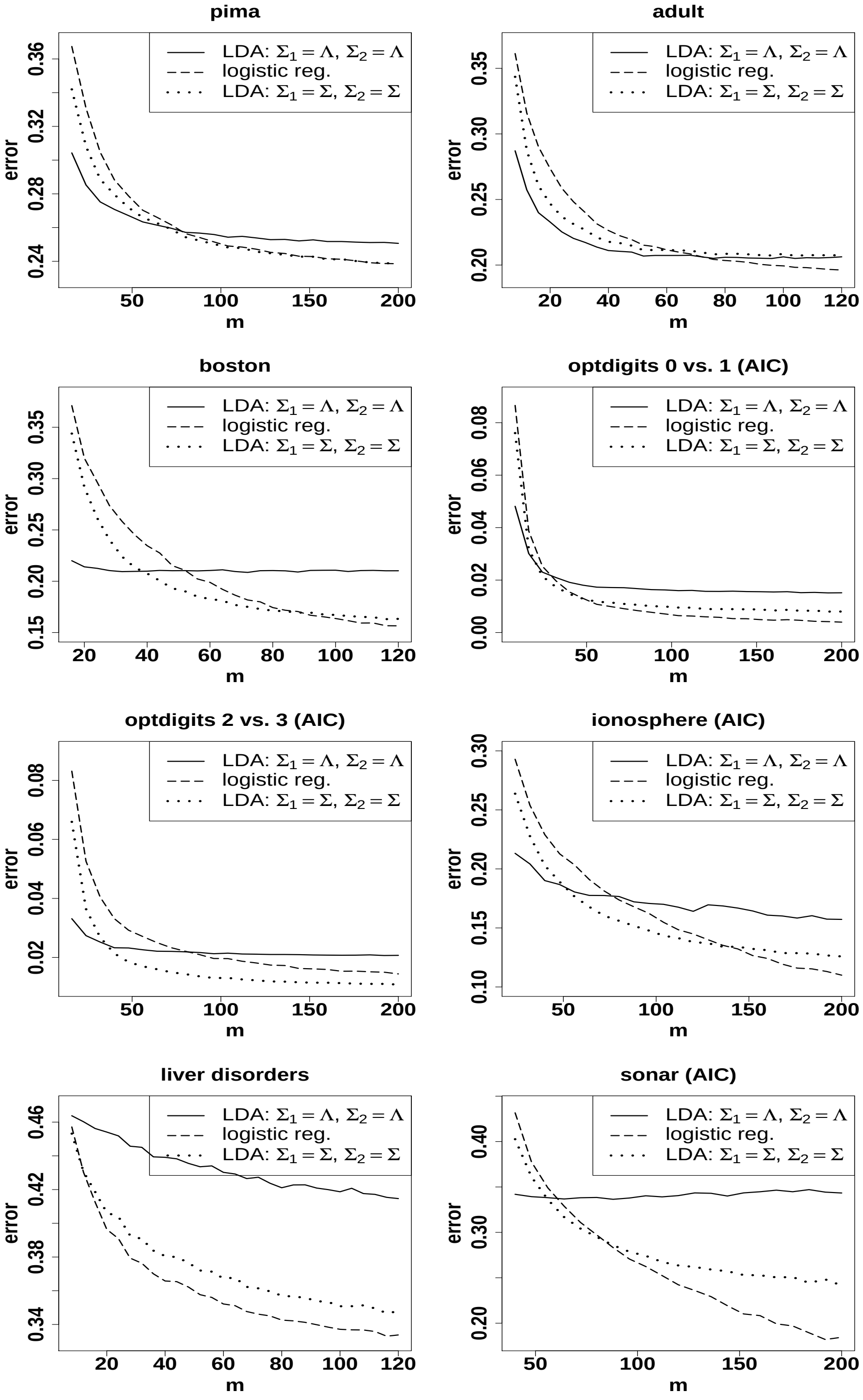

Figure 1. Plots of misclassification error rate vs. training-set size m (averaged over 1000 random training/test set splits) for the continuous UCI datasets, with regard to linear discrimination.

Secondly, the performance of LDA-Σ is worse than that of LDA-Λ when the training-set size m is small, but better than that of the latter when m is large.

Thirdly, the performance of LDA-Σ is better than that of linear logistic regression when m is small, but is more or less comparable with that of the latter when m is large.

Fourthly, pre-processing with variable selection can reveal the distinction in performance between generative and discriminative classifiers with fewer training observations.

Therefore, considering LDA-Λ vs. linear logistic regression, there is strong evidence to support the claim that the discriminative classifier has lower asymptotic error rate while the generative classifier may approach its (higher) asymptotic error rate much faster. However, considering LDA-Σ vs. linear logistic regression, the evidence is not so strong, although the claim may still be made.

3.2. Quadratic Discrimination On Continuous Datasets

As a natural extension of the comparison between LDA-Λ (with a common diagonal covariance matrix Λ across the two classes), LDA-Σ (with a common full covariance matrix Σ) and linear logistic regression that was presented in Section 3.1, this section presents the comparison between QDA-Λ_g (with two unequal diagonal covariance matrices Λ_1 and Λ_2), QDA-Σ_g (with two unequal full covariance matrices Σ_1 and Σ_2) and quadratic logistic regression.

Using the 8 continuous UCI datasets, all the settings are the same as those in Section 3.1 except for the following aspects.

First, considering that in the quadratic logistic regression model there are $p(p-1)/2$ interaction terms between the features in a p-dimensional feature space, a large number of interactions when the dimensionality p is high, the model is constrained to contain only the intercept, the p features and their p squared terms, so as to make the estimation of the model more feasible and interpretable.

Secondly, for the same reason as explained at the end of Section 1, in the reported plots of misclassification error rate vs. m without variable selection, only the results for $m > 2p$ are reliable for comparison since there are $2p$ predictor variables in the quadratic logistic regression model. Hence, only the results for $m > 2p$ are shown in Figure 2.

Thirdly, the datasets are randomly split into training sets and test sets 100 times rather than 1000 times for each sampled training-set size m because of the higher computational complexity of the quadratic models compared with that of the linear models.

In general, our study of these continuous datasets, as shown in Figure 2, suggests quite similar conclusions to those in Section 3.1, through substituting QDA-Λ_g for LDA-Λ, QDA-Σ_g for LDA-Σ, and quadratic logistic regression for linear logistic regression.

3.3. Linear Discrimination On Discrete Datasets

For the discrete datasets, as was done by [6], all the continuous features are removed and only the discrete features are used. The results are entitled ‘multinomial’ in the following figures if a dataset includes multinomial features, and otherwise are entitled ‘binomial’. Meanwhile,

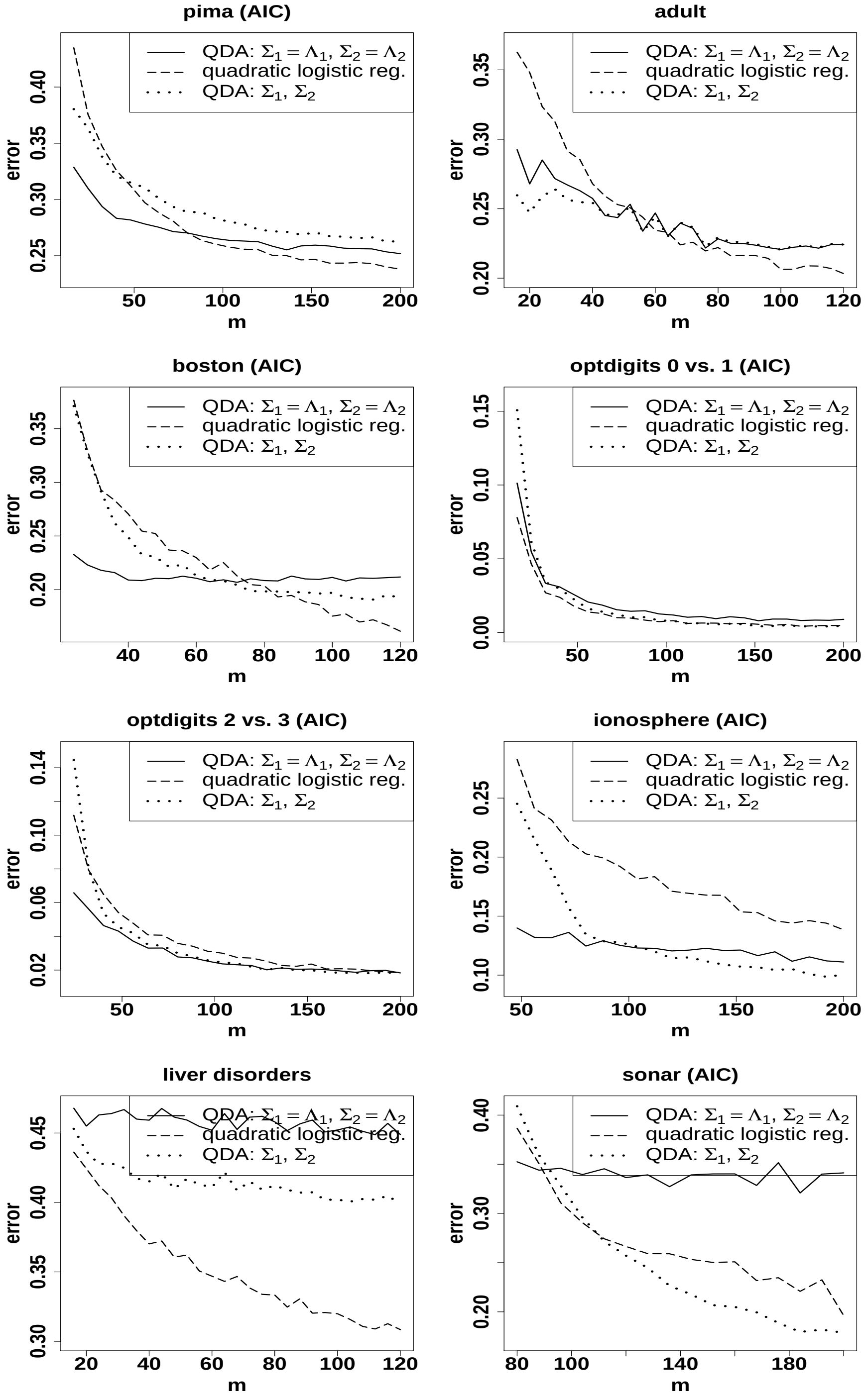

Figure 2. Plots of misclassification error rate vs. training-set size m (averaged over 100 random training/test set splits) for the continuous UCI datasets, with regard to quadratic discrimination.

any observation with missing features is removed from the datasets, as is any feature with only a single value for all the observations.

Table II. Description of discrete datasets.

Dataset	N_0	N	p	$\tilde{p}$	p_{AIC}	$\tilde{p}_{AIC}$	$\mathbf{1}_{\{2R-NB\}}$
Promoters	106	106	57	171	7	21	0
Lymphography	148	142	17	38	10	27	0
Breast cancer	286	277	9	30	4	6	0
Voting recorders	435	232	16	16	11	11	1
Lenses	24	24	4	5	1	1	0
Sick	2800	500	12	15	4	7	1
Adult	32561	1000	5	20	5	20	1

A brief description of the discrete datasets can be found in Table II, which includes the indicator $\mathbf{1}_{\{2R-NB\}} \in \{1, 0\}$ of whether or not the two regimes are observed between the naïve Bayes classifier and linear logistic regression.

Our results are shown in Figure 3 for some $m > \tilde{p}$ or $m > \tilde{p}_{AIC}$, with dummy variables taken into account for the multinomial features.

In general, our study of these discrete datasets suggests that, in the comparison of the naïve Bayes classifier vs. linear logistic regression, the pattern of our results can be said to be similar to that of [6].

4. Linear Discrimination On Simulated Datasets

In this section, 16 simulated datasets are used to compare the performance of LDA-Λ, LDA-Σ and linear logistic regression. The samples are simulated from bivariate normal distributions, bivariate Student's t-distributions, bivariate log-normal distributions and mixtures of 2

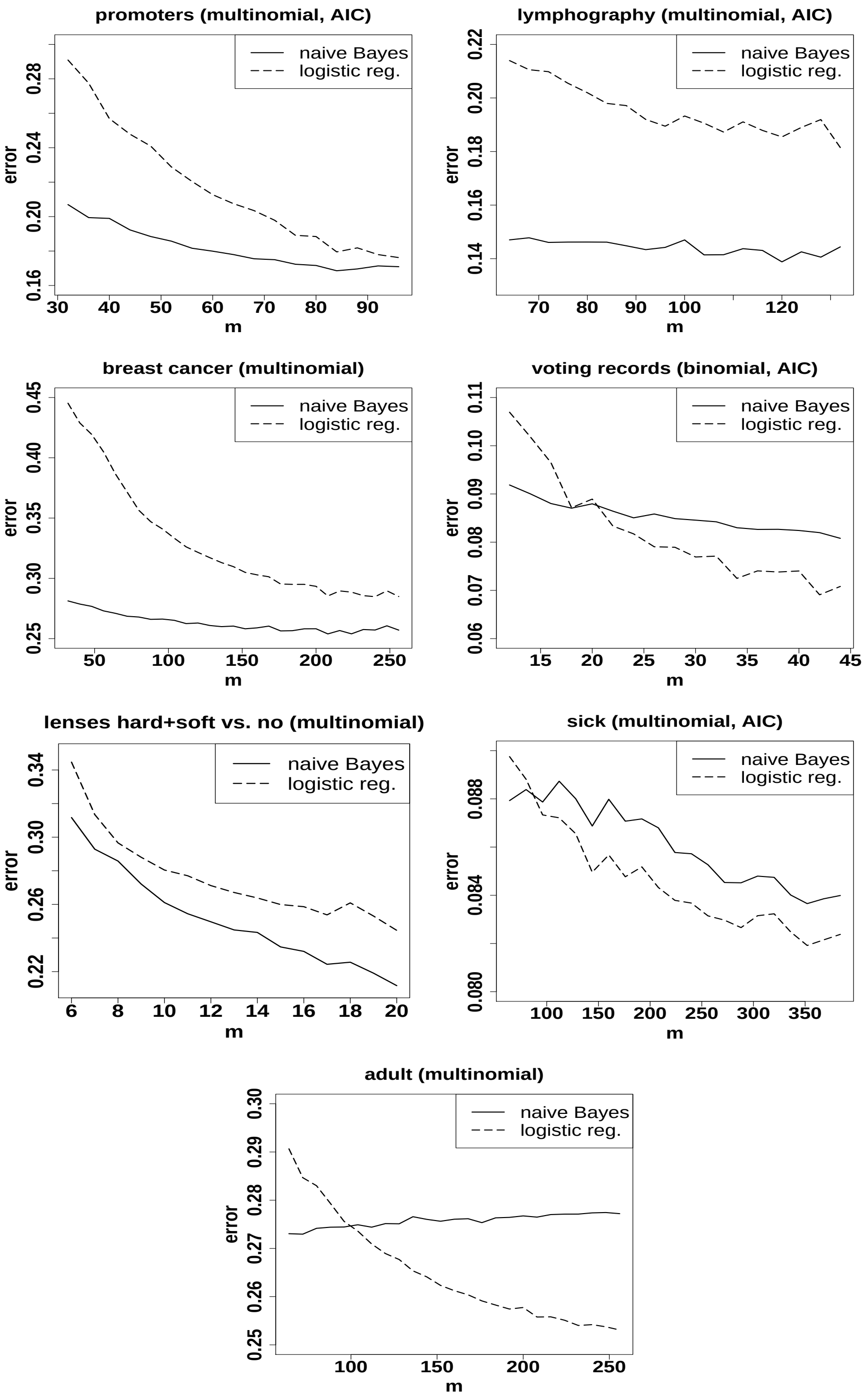

Figure 3. Plots of misclassification error rate vs. training-set size m (averaged over 1000 random training/test set splits) for the discrete UCI datasets, with regard to linear discrimination.

bivariate normal distributions, with 4 datasets for each of these 4 types of distribution. Within each dataset there are 1000 simulated samples, which are divided equally into 2 classes. The simulations from the bivariate log-normal distributions and normal mixtures are based on an R function *mvrnorm* for simulating from a multivariate normal distribution from a contributed R package **MASS**, and the simulation from the bivariate Student's t-distribution is implemented by an R function *rmvt* from a contributed R package **mvtnorm**. Differently from the UCI datasets, the simulated data are not rescaled into the range $[0, 1]$ and no variable selection is used since the feature space is only of dimension two.

4.1. Normally Distributed Data

Four simulated datasets are randomly generated from two bivariate normal distributions, $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$, where $\mu_1 = (1, 0)^T$, $\mu_2 = (-1, 0)^T$ and Σ_1 and Σ_2 are subject to four different types of constraint specified as having equal diagonal or full covariance matrices $\Sigma_1 = \Sigma_2$ and having unequal diagonal or full covariance matrices $\Sigma_1 \neq \Sigma_2$.

Similarly to what was done for the UCI datasets, for each sampled training-set size m, we perform 1000 random splits of the 1000 samples of each simulated dataset into a training set of size m and a test set of size $1000 - m$, and report the average misclassification error rates over these 1000 test sets. The training set is required to have at least 1 sample from each of the two classes. In such a way, LDA-Λ and LDA-Σ are compared with linear logistic regression, in terms of misclassification error rate, with the following results shown in Figure 4.

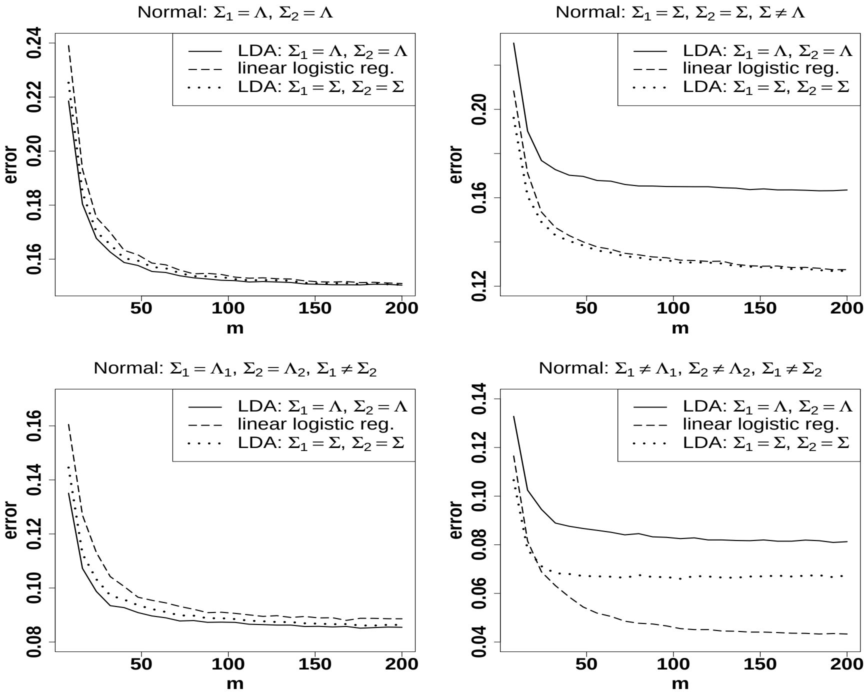

Figure 4. Plots of misclassification error rate vs. training-set size m (averaged over 1000 random training/test set splits) for simulated bivariate normally distributed data for two classes.

The dataset for the top-left panel of Figure 4 has $\Sigma_1 = \Sigma_2 = \Lambda$ with a diagonal matrix $\Lambda = \text{Diag}(1, 1)$, such that the data satisfy the assumptions underlying LDA-Λ. The dataset for the top-right panel has $\Sigma_1 = \Sigma_2 = \Sigma$ with a full matrix $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, such that the data satisfy the assumptions underlying LDA-Σ. The dataset for the bottom-left panel has $\Sigma_1 = \Lambda_1, \Sigma_2 = \Lambda_2$ with diagonal matrices $\Lambda_1 = \text{Diag}(1, 1)$ and $\Lambda_2 = \text{Diag}(0.25, 0.75)$, such that the homogeneity of the covariance matrices is violated. The dataset for the bottom-right panel has

$\Sigma_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ and $\Sigma_2 = \begin{bmatrix} 0.25 & 0.5 \\ 0.5 & 1.75 \end{bmatrix}$, such that both the homogeneity of the covariance matrices and the conditional independence (uncorrelatedness) of the features within a class are violated.

4.2. Student's t-Distributed Data

Four simulated datasets are randomly generated from two bivariate Student's t-distributions, both distributions with degrees of freedom $\nu = 3$. The values of class means μ_1 and μ_2, the four types of constraint on Σ_1 and Σ_2, and other settings of the experiments are all the same as those in Section 4.1.

The results are shown in Figure 5, where for each panel the constraint with regard to Σ_1 and Σ_2 is the same as the corresponding one in Figure 4, except for a scalar multiplier $\nu/(\nu - 2)$.

4.3. Log-normally Distributed Data

Four simulated datasets are randomly generated from two bivariate log-normal distributions, whose logarithms are normally distributed as $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$, respectively. The values of μ_1 and μ_2, the four types of constraint on Σ_1 and Σ_2, and other settings of the experiments are all the same as those in Section 4.1.

By definition, if a p-variate random vector $\mathbf{x} \sim \mathcal{N}(\mu(\mathbf{x}), \Sigma(\mathbf{x}))$, then a p-variate vector $\tilde{\mathbf{x}}$ of the exponentials of the components of $\mathbf{x}$ follows a p-variate log-normal distribution, *i.e.*, $\tilde{\mathbf{x}} = \exp(\mathbf{x}) \sim \log \mathcal{N}(\mu(\tilde{\mathbf{x}}), \Sigma(\tilde{\mathbf{x}}))$, where the i-th element $\mu^{(i)}(\tilde{\mathbf{x}})$ of the mean vector and the (i, j)-th

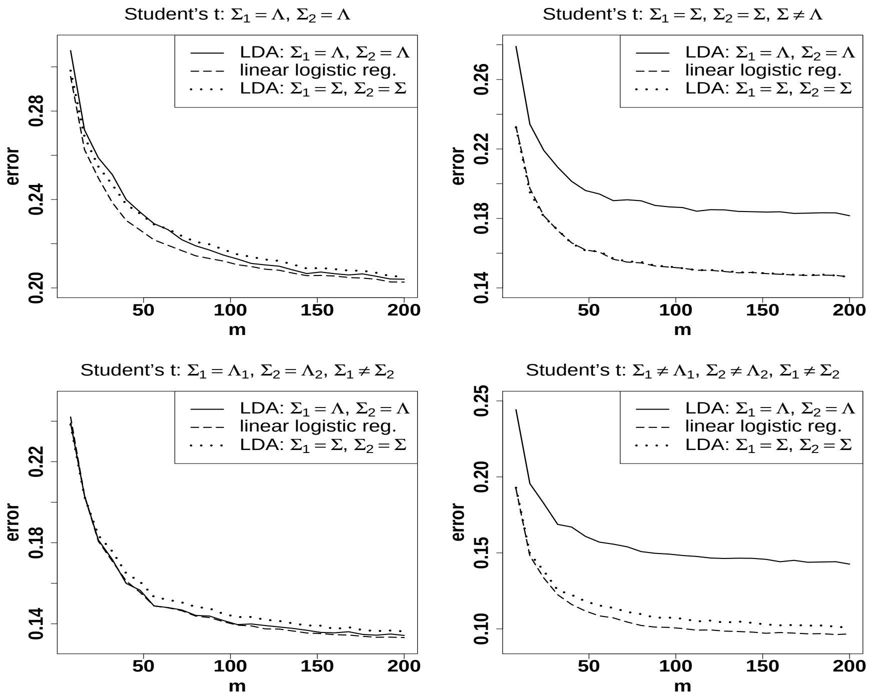

Figure 5. Plots of misclassification error rate vs. training-set size m (averaged over 1000 random training/test set splits) for simulated bivariate Student's t-distributed data for two classes.

element $\Sigma^{(i,j)}(\tilde{\mathbf{x}})$ of the covariance matrix, $i, j = 1, \ldots, p$, are

$$\mu^{(i)}(\tilde{\mathbf{x}}) = e^{\mu^{(i)}(\mathbf{x}) + \frac{\Sigma^{(i,i)}(\mathbf{x})}{2}},$$

$$\Sigma^{(i,j)}(\tilde{\mathbf{x}}) = (e^{\Sigma^{(i,j)}(\mathbf{x})} - 1)e^{\mu^{(i)}(\mathbf{x}) + \mu^{(j)}(\mathbf{x}) + \frac{\Sigma^{(i,i)}(\mathbf{x}) + \Sigma^{(j,j)}(\mathbf{x})}{2}} .$$

It follows that, if the components of its logarithm $\mathbf{x}$ are independent and normally distributed, the components of the log-normally distributed multivariate random variable $\tilde{\mathbf{x}}$ are uncorrelated. In other words, if $\mathbf{x} \sim \mathcal{N}(\mu(\mathbf{x}), \Lambda(\mathbf{x}))$, then $\tilde{\mathbf{x}} = \exp(\mathbf{x}) \sim \log\mathcal{N}(\mu(\tilde{\mathbf{x}}), \Lambda(\tilde{\mathbf{x}}))$. However, as shown by the equations above, $\Lambda(\tilde{\mathbf{x}})$ is determined by both $\mu(\mathbf{x})$ and $\Lambda(\mathbf{x})$, so that $\Sigma_1(\mathbf{x}) = \Sigma_2(\mathbf{x})$ may not mean $\Sigma_1(\tilde{\mathbf{x}}) = \Sigma_2(\tilde{\mathbf{x}})$.

Therefore, if we consider in our cases $\mu_1 \neq \mu_2$, it can be expected that the pattern of performance of the classifiers for the datasets with equal covariance matrices $\Sigma_1 = \Sigma_2$ in the underlying normal distributions could be similar to that for the datasets with unequal covariance matrices $\Sigma_1 \neq \Sigma_2$, since in both cases the covariance matrices of the log-normally distributed variables are in fact unequal. In this context, it makes more sense to compare the classifiers in situations with diagonal and full covariance matrices of the underlying normally distributed data, respectively, rather than those with equal and unequal covariance matrices.

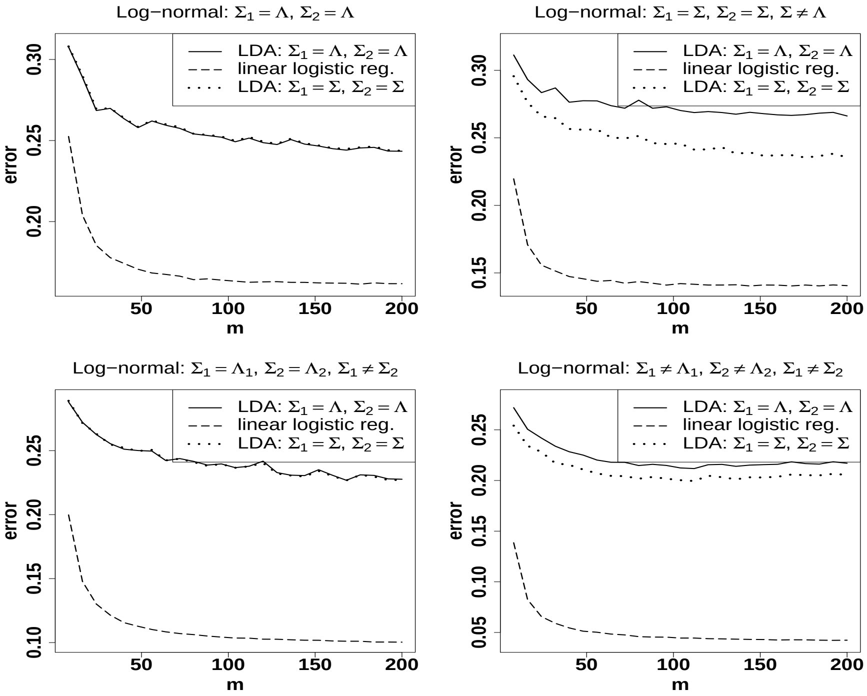

Figure 6. Plots of misclassification error rate vs. training-set size m (averaged over 1000 random training/test set splits) for simulated bivariate log-normally distributed data for two classes.

The results are shown in Figure 6, where for each panel the constraint with regard to Σ_1 and Σ_2 is the same as the corresponding one in Figure 4.

4.4. Normal Mixture Data

Compared with the normal distribution, the Student's t-distribution and the log-normal distribution used in Sections 4.1, 4.2 and 4.3 for the comparison of the classifiers, the mixture of normal distributions is a better approximation to real data in a variety of situations. In this section, 4 simulated datasets, each consisting of 1000 samples, are randomly generated from two mixtures, each of two bivariate normal distributions, with 250 samples from each mixture component. The two components, A and B, of the mixture for Class 1 are normally distributed with distributions $\mathcal{N}(\mu_{1A}, \Sigma_1)$ and $\mathcal{N}(\mu_{1B}, \Sigma_1)$, respectively, where $\mu_{1A} = (1, 0)^T$ and $\mu_{1B} = (3, 0)^T$; and the two components, C and D, of the mixture for Class 2 are normally distributed with probability density functions $\mathcal{N}(\mu_{2C}, \Sigma_2)$ and $\mathcal{N}(\mu_{2D}, \Sigma_2)$, respectively, where $\mu_{2C} = (-1, 0)^T$ and $\mu_{2D} = (-3, 0)^T$. In such a way, when Σ_1 and Σ_2 are subject to the four different types of constraint with regard to Σ_1 and Σ_2 as previously discussed, the covariance matrices of the two mixtures will be subject to the same constraints. Other settings of the experiments are all the same as that in Section 4.1.

The results are shown in Figure 7, where for each panel the constraint with regard to Σ_1 and Σ_2 is the same as the corresponding one in Figure 4.

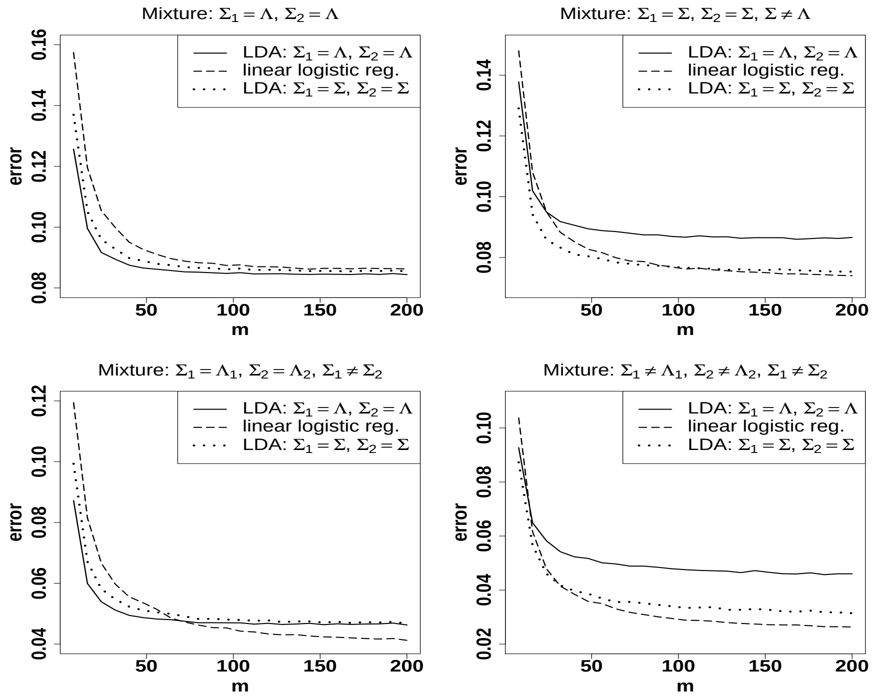

Figure 7. Plots of misclassification error rate vs. training-set size m (averaged over 1000 random training/test set splits) for simulated bivariate 2-component normal mixture data for two classes.

4.5. Summary of Linear Discrimination for Simulated Datasets

In general, our study of these simulated continuous datasets suggests the following conclusions.

First, when the data are consistent with the assumptions underlying LDA-Λ or LDA-Σ, as shown in the top-left and top-right panels of Figure 4, both methods can perform the best among them and linear logistic regression, throughout the range of the training-set size m in our study. In these cases, there is no evidence to support the claim that the discriminative classifier has lower asymptotic error rate while the generative classifier may approach its (higher) asymptotic error rate much faster.

Secondly, when the data violate the assumptions underlying the LDAs, linear logistic regression generally performs better than the LDAs, in particular when m is large. This pattern is especially clear, as shown in Figure 6 for the log-normally distributed data, the distributions of which are heavy-tailed, asymmetric and thus in some sense less 'Gaussian' than Student's t and normal-mixture distributions in our experiments. In this case, there is strong evidence to support the claim that the discriminative classifier has lower asymptotic error rate, but there is no convincing evidence to support the claim that the generative classifier may approach its (higher) asymptotic error rate much faster.

Finally, when the covariance matrices are non-diagonal, LDA-Σ performs remarkably better than LDA-Λ and more remarkably when m is large; when the covariance matrices are diagonal, LDA-Λ performs generally better than LDA-Σ and more so when m is large.

5. Comments on the Two Regimes of Performance regarding Training-Set Size

Based on the theoretical analysis and empirical comparison between LDA-Λ or the naïve Bayes classifiers and linear logistic regression, Ref. [6] claim that there are two distinct regimes of performance with regard to the training-set size m. However, our empirical results, as shown in Tables I and II, could not convincingly support the claim. Furthermore, our simulation studies, as presented in Section 4, failed to find the two regimes when the data either conformed to the assumptions underlying the generative classifiers, as shown in Figure 4, or heavily violated the assumptions, as shown in Figure 6.

Therefore, besides commenting on the pairing of the compared classifiers in Section 2.2, we shall clarify the claim further through commenting on the reliability of the two regimes.

Suppose we have a training set $\{(y_{tr}^{(i)}, \mathbf{x}_{tr}^{(i)})\}_{i=1}^{m}$ of m independent observations and a test set $\{(y_{te}^{(i)}, \mathbf{x}_{te}^{(i)})\}_{i=1}^{N-m}$ of $N-m$ independent observations, where $\mathbf{x}^{(i)} = (x_1^{(i)}, \ldots, x_p^{(i)})^T$ is the i-th observed p-variate feature vector $\mathbf{x}$, and $y^{(i)} \in \{1, 2\}$ is its observed univariate class label. Let us also assume that each observation $\{(y^{(i)}, \mathbf{x}^{(i)})\}$ follows an identical distribution so that testing based on the training results makes sense. In order to simplify the notation, let $\underline{\mathbf{x}}_{tr}$ denote $\{(\mathbf{x}_{tr}^{(i)})\}_{i=1}^{m}$, and similarly define $\underline{\mathbf{x}}_{te}$, $\underline{y}_{tr}$ and $\underline{y}_{te}$. Meanwhile, a discriminant function $\lambda(\alpha) = \log\{p(y=1|\mathbf{x})/p(y=2|\mathbf{x})\}$, which is equivalent to a Bayes classifier $\hat{y}(\mathbf{x}) = \mathrm{argmax}_y\, p(y|\mathbf{x})$, is used for the 2-class classification.

5.1. For Discriminative Classifiers

Discriminative classifiers estimate the parameter α of the discriminant function $\lambda(\alpha)$ through $\hat{\alpha} = \mathrm{argmax}_\alpha\, p(\underline{y}_{tr}|\underline{\mathbf{x}}_{tr}, \alpha)$, the maximisation of a conditional probability; such an estimation procedure can be regarded as a kind of maximum likelihood estimation with $p(\underline{y}_{tr}|\underline{\mathbf{x}}_{tr}, \alpha)$ as the likelihood function. It is well known that, if the $0-1$ loss function is used so that the misclassification error rate is the total risk, the Bayes classifiers will attain the minimum error rate [9]. This implies that, under such a loss function, the discriminative classifiers are in fact using the same criterion to optimise the estimation of the parameter α and the performance of classification.

In this context, the following claims, supported by the simulation study in Section 4, can be proposed.

First, if the same dataset is used to train and test, *i.e.*, $\underline{\mathbf{x}}_{tr}$ as $\underline{\mathbf{x}}_{te}$ and $\underline{y}_{tr}$ as $\underline{y}_{te}$, then the discriminative classifiers should always provide the best performance, no matter how large the training-set size m is, provided that the $0 - 1$ loss function is used and the modelling of $p(y|\mathbf{x}, \alpha)$, such as the linearity of $\lambda(\alpha)$, is correctly specified for all the observations, and thus the only work that remains is to estimate accurately the parameter α.

Secondly, if m is large enough to make $(\underline{y}_{tr}, \underline{\mathbf{x}}_{tr})$ representative of all the observations including $(\underline{y}_{te}, \underline{\mathbf{x}}_{te})$, then the discriminative classifiers should also provide the best prediction performance on $(\underline{y}_{te}, \underline{\mathbf{x}}_{te})$, *i.e.*, with the best asymptotic performance, provided that the modelling of $p(y|\mathbf{x}, \alpha)$ is correctly specified for all the observations.

Finally, if m is not large enough to make $(\underline{y}_{tr}, \underline{\mathbf{x}}_{tr})$ representative of all the observations, and $(\underline{y}_{te}, \underline{\mathbf{x}}_{te})$ is not exactly the same as $(\underline{y}_{tr}, \underline{\mathbf{x}}_{tr})$, then the discriminative classifiers may not necessarily provide the best prediction performance on $(\underline{y}_{te}, \underline{\mathbf{x}}_{te})$, even though the modelling of $p(y|\mathbf{x}, \alpha)$ may be correct.

5.2. For Generative Classifiers

Generative classifiers estimate the parameter α of the discriminant function $\lambda(\alpha)$ through first maximising a joint probability function, i.e. $\hat{\theta} = \text{argmax}_{\theta}\, p(\underline{y}_{tr}, \underline{\mathbf{x}}_{tr}|\theta)$, to obtain a maximum likelihood estimate

(MLE) $\hat{\theta}$ of θ, the parameter of the joint distribution of $(y, \mathbf{x})$, and then calculate $\hat{\alpha}$ as a function $\alpha(\theta)$ at $\hat{\theta}$. Under some regularity conditions, such as the existence of the first and second derivatives of the log-likelihood function and the inverse of the Fisher information matrix $I(\theta)$, the MLE $\hat{\theta}$ is asymptotically unbiased, efficient and normally distributed. Accordingly, by the delta method, $\hat{\alpha}$ is also asymptotically normally distributed, unbiased and efficient, given the existence of the first derivative of the function $\alpha(\theta)$.

Therefore, the following claims, supported by the simulation study in Section 4, can be proposed.

First, asymptotically, the generative classifiers will provide the best prediction performance on $(\underline{y}_{te}, \underline{\mathbf{x}}_{te})$, dependent on the premise that the modelling of $p(y, \mathbf{x}|\theta)$, instead of $p(y|\mathbf{x}, \alpha)$, is correctly specified for all the observations.

Secondly, if m is large enough to make $(\underline{y}_{tr}, \underline{\mathbf{x}}_{tr})$ representative of all the observations including $(\underline{y}_{te}, \underline{\mathbf{x}}_{te})$, then the generative classifiers should also provide the best prediction performance on $(\underline{y}_{te}, \underline{\mathbf{x}}_{te})$, *i.e.*, with the best asymptotic performance, given that the modelling of $p(y, \mathbf{x}|\theta)$, instead of $p(y|\mathbf{x}, \alpha)$, is correctly specified for all the observations.

Finally, if m is not large enough to make $(\underline{y}_{tr}, \underline{\mathbf{x}}_{tr})$ representative of all the observations, then the generative classifiers may not necessarily provide the best prediction performance on $(\underline{y}_{te}, \underline{\mathbf{x}}_{te})$.

5.3. Summary

In summary, it may not be so reliable to claim the existence of the two distinct regimes of performance between the generative and discriminative classifiers with regard to the training-set size m.

For real world datasets such as those demonstrated in Sections 3.1 and 3.3, so far there is no theoretically correct, general criterion for choosing between the discriminative and the generative classifiers; the choice depends on the relative confidence we have in the correctness of the specification of either $p(y|\mathbf{x})$ or $p(y, \mathbf{x})$ for the data. This can be to some extent a demonstration of why Ref. [3] and [7] prefer LDA when no model mis-specification occurs but other empirical studies may prefer linear logistic regression instead.

Ref. [6] provided theoretical proof that the discriminative classifiers need $m \in \Omega(p)$ (i.e., $m \geq M_1 \, p$ where $M_1 > 0$) training observations to approach its asymptotic error rate with high probability, whereas the generative classifiers need only $m \in \Omega(\log(p))$ (i.e., $m \geq M_2 \log(p)$ where $M_2 > 0$) training observations. We observe the following. First, for two distinct regimes to occur, it is necessary that $M_1 \, p \geq M_2 \log(p)$. Secondly, such a higher efficiency of the generative classifiers might be also attained because of the bias induced by its model mis-specification, such as using LDA-Λ/the naïve Bayes classifiers for the cases in which it would be better to adopt LDA-Σ/QDA-Σ_g. For real-world data, application of such a mis-specified model is likely; the bias-variance tradeoff may then play a role in determining the occurrence of the two distinct regimes.

In addition, a similar pattern of two distinct regimes with regard to m was also reported in Ref. [8], based on the performance of logistic regression and tree induction; they found that logistic regression performs better with smaller m and tree induction with larger m. Therefore, although tree induction and logistic regression are not a pair of generative and discriminative classifiers, it could be interesting to explore such a pattern for other pairs of classifiers.

Acknowledgements

The authors thank Andrew Y. Ng for communication about the implementation of the empirical studies in this paper. We are grateful to the associate editor, anonymous referees and David J. Hand for advice about the structure, references, experimental illustration and interpretation of this manuscript. The work also benefited from our participation in the Research Programme on 'Statistical Theory and Methods for Complex, High-Dimensional Data' at the Isaac Newton Institute for Mathematical Sciences in Cambridge.

References

1. Asuncion, A. and D. J. Newman: 2007, 'UCI Machine Learning Repository'. Irvine, CA: University of California, School of Information and Computer Science. http://www.ics.uci.edu/∼mlearn/MLRepository.html.
2. Dawid, A. P.: 1976, 'Properties of diagnostic data distributions'. *Biometrics* **32**(3), 647–658.

3. Efron, B.: 1975, 'The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis'. *Journal of the American Statistical Association* **70**(352), 892–898.
4. Hand, D. J.: 2006, 'Classifier technology and illusion of progress (with discussion)'. *Statistical Science* **21**, 1–34.
5. Lim, T.-S. and W.-Y. Loh: 1996, 'A comparison of tests of equality of variances'. *Computational Statistics & Data Analysis* **22**(3), 287–301.
6. Ng, A. Y. and M. I. Jordan: 2001, 'On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes'. In: *NIPS*. pp. 841–848.
7. O'Neill, T. J.: 1980, 'The general distribution of the error rate of a classification procedure with application to logistic regression discrimination'. *Journal of the American Statistical Association* **75**(369), 154–160.
8. Perlich, C., F. Provost, and J. S. Simonoff: 2003, 'Tree induction vs. logistic regression: A learning-curve analysis'. *Journal of Machine Learning Research* **4**(211–255).
9. Ripley, B. D.: 1996, *Pattern Recognition and Neural Networks*. New York: Cambridge University Press.
10. Rubinstein, Y. D. and T. Hastie: 1997, 'Discriminative vs. informative learning'. In: *KDD*. pp. 49–53.
11. Shapiro, S. S. and M. B. Wilk: 1965, 'An analysis of variance test for normality (complete samples)'. *Biometrika* **52**(3-4), 591–611.
12. Titterington, D. M., G. D. Murray, L. S. Murray, D. J. Spiegelhalter, A. M. Skene, J. D. F. Habbema, and G. J. Gelpke: 1981, 'Comparison of Discrimination Techniques Applied to a Complex Data Set of Head Injured Patients (with discussion)'. *Journal of the Royal Statistical Society. Series A (General)* **144**(2), 145–175.
13. Verboven, S. and M. Hubert: 2005, 'LIBRA: a MATLAB Library for Robust Analysis'. *Chemometrics and Intelligent Laboratory Systems* **75**(2), 127–136.