Aspects of Generative and Discriminative Classifiers

Jing-Hao Xue

A Dissertation Submitted to the
University of Glasgow
for the degree of
Doctor of Philosophy

Department of Statistics, University of Glasgow

June 2008

Summary

In recent years, under the new terminology of generative and discriminative classifiers, research interest in classical statistical approaches to discriminant analysis has re-emerged in the machine learning community. In discriminant analysis, observations with features $\mathbf{x}$ measured are classified into classes labelled by a categorical variable y. *Generative classifiers*, also termed the sampling paradigm, 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 $p(\mathbf{x}|y)p(y)$, where $p(\mathbf{x}|y)$ is a data-generating process (DGP), and learn the model parameters through maximisation of the likelihood with respect to $p(\mathbf{x}|y)p(y)$. *Discriminative classifiers*, also termed the diagnostic paradigm, 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 $p(y|\mathbf{x})$.

In order to exploit the best of both worlds, it is necessary to first compare generative and discriminative classifiers and then combine them. In this thesis, we first performed some empirical and simulation studies to provide extension of and make comments on a highly-cited report (Ng and Jordan, 2001), which compared the naïve Bayes classifier or normal-based linear discriminant analysis (LDA) with linear logistic regression (LLR). Then we studied extensively two hybrid-learning techniques, namely the hybrid generative-discriminative algorithm (Raina et al., 2003) and the generative-discriminative tradeoff (GDT) approach (Bouchard and Triggs, 2004), for combining the generative and discriminative classifiers. Based on our results from these studies, we proposed a joint generative-discriminative modelling approach to classification. In addition, we extended our investigation to generative and discriminative hidden Markov models, the latent variable models for structured data. We also developed discriminative approaches for a specific application, that of histogram-based image thresholding.

The contributions of this thesis are the following.

First, Ng and Jordan (2001) claimed that there exist two distinct regimes of performance between the generative and discriminative classifiers with regard to the training-set size; however, our empirical and simulation studies, as presented in Chapter 2, suggest that it is not so reliable to claim such an existence of the two distinct regimes. 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 you have in the correctness of the specification of either $p(y|\mathbf{x})$ or $p(\mathbf{x}, y)$. This can be to some extent a demonstration of why Efron (1975) and O'Neill (1980) prefer LDA but other empirical studies may prefer LLR instead. Furthermore, we suggest that pairing of either LDA assuming a common diagonal covariance matrix (LDA-Λ) or the naïve Bayes classifier and LLR 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 LLR to be generalised to all generative and discriminative classifiers.

Secondly, in Chapter 3, we present the interpretation and asymptotic relative efficiency (ARE) of the GDT approach for linear and quadratic normal discrimination without model mis-specification, and compare its ARE with those of its generative and discriminative counterparts. The classification performance of the GDT is compared with those of LDA and LLR on simulated datasets. We argue that the GDT is a generative model integrating both discriminative and generative learning. It is therefore sensitive to model mis-specification of the data-generating process and, in practice, its discriminative component may behave differently from a truly discriminative approach. Amongst the three approaches that we compare, the asymptotic efficiency of the GDT is lower than that of the generative approach when no model mis-specification occurs. In addition, without model mis-specification, LDA performs the best; with model mis-specification, the GDT may perform the best at an optimal tradeoff between its discriminative and generative components, and LLR, a truly discriminative classifier, in general performs well when the training-sample size is reasonably large.

Thirdly, in Chapter 4, we interpret the hybrid algorithm from three perspectives, namely class-conditional probabilities, class-posterior probabilities and loss functions underlying the model. We suggest that the hybrid algorithm is by nature a generative model with its parameters learnt through both generative and discriminative approaches, in the sense that it assumes a scaled data-generation process and uses scaled class-posterior probabilities to perform discrimination. Our suggestion can also be applied to its multi-class extension. In addition, using

simulated and real-world data, we compare the performance of the normalised hybrid algorithm as a classifier with that of the naïve Bayes classifier and LLR. Our simulation studies suggest in general the following: if the covariance matrices are diagonal matrices, the naïve Bayes classifier performs the best; if the covariance matrices are full matrices, LLR performs the best. Our studies also suggest that the hybrid algorithm may provide worse performance than either the naïve Bayes classifier or LLR alone.

Fourthly, based on our studies presented in Chapters 2, 3 and 4, we propose in Chapter 5 a joint generative-discriminative modelling (JGD) approach to classification, by partitioning variables into two subsets based on statistical tests of the DGP. Our JGD approach adopts statistical tests, such as normality tests, of the assumed DGP for each variable to justify the use of generative approaches for the variables which satisfy the tests and of discriminative approaches for other variables. Such a partition of variables and a combination of generative and discriminative approaches are derived in a probabilistic rather than a heuristic way. We have concentrated on particular choices for the generative and discriminative components of our models, but the overall principle is quite general and can accommodate many other special versions. Of course, we must ensure that the assumptions underlying the resulting generative classifiers can be tested statistically. Numerical results from real UCI and gene-expression data and from simulated data demonstrate promising performance of this new approach for practical application to both low- and high-dimensional data.

Fifthly, in Chapter 6, we study the assumption of "mutual information independence", which is used by Zhou (2005) for deriving the so-called discriminative hidden Markov model (D-HMM). We suggest that the mutual information assumption (6.6) results in the D-HMM, while another mutual information assumption (6.12) results in its generative counterpart, the G-HMM. However, in practice, whether or not the assumptions are reasonable and how the corresponding HMMs perform can be data-dependent; research efforts to explore an adaptive switching between or combination of these two models may be worthwhile. Meanwhile, we suggest that the so-called output-dependent HMMs could be represented in a state-dependent manner, and vice versa, essentially by application of Bayes' theorem.

Finally, in Chapter 7, we present discriminative approaches to histogram-based image thresholding, in which the optimal threshold is derived from the maximum likelihood based on the conditional distribution $p(y|x)$ of y, the class indicator of a grey level x, given x. The discriminative approaches can be regarded as discriminative extensions of the traditional gen-

erative approaches to thresholding, such as Otsu's method (Otsu, 1979) and Kittler and Illingworth's minimum error thresholding (MET) (Kittler and Illingworth, 1986). As illustrations, we develop discriminative versions of Otsu's method and MET by using discriminant functions corresponding to the original methods to represent $p(y|x)$. These two discriminative thresholding approaches are compared with their original counterparts on selecting thresholds for a variety of histograms of mixture distributions. Results show that the discriminative Otsu method consistently provides relatively good performance. Although being of higher computational complexity than the original methods in parameter estimation, its robustness and model simplicity can justify the discriminative Otsu method for scenarios in which the risk of model mis-specification is high and the computation is not demanding.

Acknowledgements

I am grateful to Professor Mike Titterington for his all-around supervision of my work for this thesis. I am in debt to my wife and my parents for their unconditional support to my PhD study in Glasgow.

My work for this thesis was supported by the award of a *Hutchison Whampoa-EPSRC Dorothy Hodgkin Postgraduate Award*; it also benefited from my 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.

Thanks to all my colleagues in the Department of Statistics for their help during my stay in Glasgow.

Thanks also to Guillaume Bouchard for constructive communications about Chapter 3 and to Andrew Y. Ng for communication about the implementation of the empirical studies in Chapter 2.

Glossary of Abbreviations

AIC: Akaike Information Criterion

AER: Asymptotic Error Rate

ARE: Asymptotic Relative Efficiency

BFGS: Broyden-Fletcher-Goldfarb-Shanno algorithm

Criterion-H: classification Criterion corresponding to the normalised Hybrid algorithm

DGP: Data-Generating Process

D-HMM: Discriminative Hidden Markov Model

ER: misclassification Error Rate

GAM: Generalised Additive Model

GDT: Generative-Discriminative Tradeoff method

G-HMM: Generative Hidden Markov Model

HMM: Hidden Markov Model

HMMSDO: Hidden Markov Models with States Depending on Observations

IRLS: Iteratively Reweighted Least Squares algorithm (also known as IWLS, or the Fisher scoring algorithm)

JDG: Joint Generative-Discriminative modelling

LDA: normal-based Linear Discriminant Analysis

LDA-Λ: LDA with a common diagonal covariance matrix

LDA-Σ: LDA with a common full covariance matrix

LL: Logarithmic Loss

LLR: Linear Logistic Regression

MAP: Maximum A Posteriori

MET: Kittler and Illingworth's Minimum Error Thresholding

MLE: Maximum Likelihood Estimate

NBC: Naïve Bayes Classifier

QDA: normal-based Quadratic Discriminant Analysis

QDA-Λ_g: QDA with unequal diagonal covariance matrices

QDA-Σ_g: QDA with unequal full covariance matrices

rpart: recursive partitioning and regression trees

DAG: Directed Acyclic Graph

d_O**:** threshold obtained from discriminative Otsu method

t_O**:** threshold obtained from Otsu's method

d_M**:** threshold obtained from discriminative MET

t_M**:** threshold obtained from MET

Contents

List of Figures

List of Tables

Chapter 1

Introduction to Generative and Discriminative Classifiers

1.1 Generative and Discriminative Classifiers

1.1.1 Definitions

In discriminant analysis, observations with measured features $\mathbf{x}$ are classified into classes labelled by a categorical variable y. The most commonly adopted discriminant rule is the maximum a posteriori (MAP) criterion: for a given observation $\mathbf{x}$, the allocated class is $\hat{y} = \operatorname{argmax}_y p(y|\mathbf{x};\alpha)$, where $\mathbf{x}$ is in general a p-variate random vector and α denotes a column vector of the parameters of the conditional distribution $p(y|\mathbf{x})$. In practice, α is unknown but can be estimated from a training set of n labelled observations $(\mathbf{x}_{1:n}, y_{1:n}) = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$.

Dawid (1976) divided the statistical modelling and learning (or parameter estimation) approaches to discrimination into two paradigms, namely, the sampling paradigm and the diagnostic paradigm. In recent years, these have re-emerged in the machine learning community under the new terminology of generative (informative) and discriminative approaches, respectively (Rubinstein and Hastie, 1997; Ng and Jordan, 2001; Raina et al., 2003; Bouchard and Triggs, 2004; McCallum et al., 2006; Bishop and Lasserre, 2007; Bouchard, 2007).

The discriminative approaches (or the approaches corresponding to the diagnostic paradigm) model $p(y_{1:n}|\mathbf{x}_{1:n};\alpha)$, without modelling the so-called data-generating process (DGP) $p(\mathbf{x}|y;\theta_{\mathrm{g}})$, where θ_{g} is the parameter vector of $p(\mathbf{x}|y)$; α is then estimated through maximisation of the conditional likelihood, *i.e.*, $\hat{\alpha} = \operatorname{argmax}_\alpha p(y_{1:n}|\mathbf{x}_{1:n};\alpha)$, which is in practice further simpli-

fied by the assumption of certain conditional-independence structure such that $p(y_{1:n}|\mathbf{x}_{1:n};\alpha) = \prod_{i=1}^{n} p(y_i|\mathbf{x}_i;\alpha)$. Thus only $p(y|\mathbf{x},\alpha)$ needs to be modelled. Hereafter, we refer to such a model and learning procedure as a *discriminative model* and *discriminative learning*, respectively. A typical discriminative classifier is logistic regression.

The generative approaches (or the approaches corresponding to the sampling paradigm) model $p(y_{1:n}|\pi)$ and $p(\mathbf{x}_{1:n}|y_{1:n};\theta_{\mathrm{g}})$, where π is the parameter vector of $p(y)$. Then, in general, $\theta = (\pi^T, \theta_{\mathrm{g}}^T)^T$ is estimated through maximum likelihood, *i.e.*, $\hat{\theta} = \mathrm{argmax}_{\theta}\, p(\mathbf{x}_{1:n}, y_{1:n};\theta)$, which is in practice further simplified by assuming certain conditional-independence structure such that $p(\mathbf{x}_{1:n}, y_{1:n};\theta) = \prod_{i=1}^{n} p(\mathbf{x}_i, y_i;\theta)$. Thus only $p(y|\pi)$ and $p(\mathbf{x}|y;\theta_{\mathrm{g}})$ need to be modelled. Hereafter, we refer to such a model and learning procedure as a *generative model* and *generative learning*, respectively. Typical generative classifiers include normal-based discriminant analysis and the naïve Bayes classifier.

As concisely characterised by Rubinstein and Hastie (1997), the generative classifiers learn the class densities, while the discriminative classifiers learn the class boundaries (*i.e.*, $p(y|\mathbf{x},\alpha)$ in our setting) without regard to the underlying class densities.

From Bayes' Theorem, which gives

$$p(y|\mathbf{x};\alpha) = \frac{p(y|\pi)p(\mathbf{x}|y;\theta_{\mathrm{g}})}{\int_y p(y|\pi)p(\mathbf{x}|y;\theta_{\mathrm{g}})} ,$$

two observations can be made. First, there is a mapping $\alpha(\theta)$ between θ and α such that the generative approaches can lead to $\hat{\alpha}$, and thereby provide working classifiers for discrimination. Secondly, the generative model is more informative than the corresponding discriminative model, and thus discriminative learning techniques can be used with a generative model. The first observation is a basic characteristic of classical generative classifiers, and the second has led to increasing research interest recently (Rubinstein, 1998; Raina et al., 2003; Bouchard and Triggs, 2004; McCallum et al., 2006).

1.1.2 Discriminant Functions

This thesis will focus on two-class discriminant analysis, where y is a binary variable. Suppose a population $\mathcal{C}$ contains two sub-populations $\mathcal{C}_1$ (with $y = 1$) and $\mathcal{C}_0$ (with $y = 0$), with respective proportions π_1 and $\pi_0 = 1 - \pi_1$; the existence of these two sub-populations requires $\pi_1 \in (0, 1)$, an open interval. In addition, the training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ contains n randomly, independently collected and labelled individuals from $\mathcal{C}$.

In the sense of minimum classification error rate, an optimal discriminant function for classifying a new individual $\mathbf{x}$ into either $\mathcal{C}_1$ or $\mathcal{C}_0$ is $g(\mathbf{x}, \alpha) = \log \frac{p(y=1|\mathbf{x})}{p(y=0|\mathbf{x})}$, the logarithm of the ratio of two posterior probabilities of the sub-population indicator y given the observed feature vector $\mathbf{x}$; *i.e.*, the new individual will be classified into $\mathcal{C}_1$ if $g(\mathbf{x}, \alpha) > 0$.

The most widely used discriminant functions are the linear discriminant function, $g(\mathbf{x}, \alpha) = \beta_0 + \beta^T \mathbf{x}$, where β_0 is a scalar, β is a p-dimensional parameter vector, $\alpha^T = (\beta_0, \beta^T)$ and $\mathbf{x}^T = (x^{(1)}, \ldots, x^{(p)})$, and the quadratic discriminant function, $g(\mathbf{x}, \alpha) = \beta_0 + \beta^T \mathbf{x} + \mathbf{x}^T \Gamma \mathbf{x}$, where Γ is a p-by-p matrix (usually symmetric) and $\alpha^T = (\beta_0, \beta^T, (\text{vech}(\Gamma))^T)$. The notation $\text{vech}(\Gamma)$ indicates a vector of distinct elements of the matrix Γ. If Γ is diagonal with diagonal components $\{\gamma_{i,i}\}_{i=1}^p$, then $g(\mathbf{x}, \alpha) = \beta_0 + \beta^T \mathbf{x} + \sum_{i=1}^p \gamma_{i,i} (x^{(i)})^2$.

The training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is used to learn the parameters α of $g(\mathbf{x}, \alpha)$. In general, the learning is performed by either discriminative approaches or generative approaches.

1.1.3 Discriminative Learning

From the definition of the discriminant function, it follows that

$$p(\mathcal{C}_1|\mathbf{x}) = p(y=1|\mathbf{x}) = \frac{e^{g(\mathbf{x},\alpha)}}{1+e^{g(\mathbf{x},\alpha)}} \ , \ p(\mathcal{C}_0|\mathbf{x}) = p(y=0|\mathbf{x}) = 1 - p(\mathcal{C}_1|\mathbf{x}) \ , \tag{1.1}$$

so that the likelihood $\mathcal{L}$ and the log-likelihood ℓ based on $p(y|\mathbf{x})$ are, respectively,

$$\mathcal{L}_{\text{d}}(\alpha) = \prod_{i=1}^n p(y_i|\mathbf{x}_i) = \prod_{i=1}^n \frac{e^{g(\mathbf{x}_i,\alpha) y_i}}{1+e^{g(\mathbf{x}_i,\alpha)}},$$

$$\ell_{\text{d}}(\alpha) = \log \mathcal{L}_{\text{d}}(\alpha) = \sum_{i=1}^n g(\mathbf{x}_i, \alpha) y_i - \sum_{i=1}^n \log(1 + e^{g(\mathbf{x}_i,\alpha)}).$$

Asymptotic theory suggests that maximisation of $\ell_{\text{d}}(\alpha)$, with respect to α, leads to an estimator $\hat{\alpha}$ of α such that the distribution of $\sqrt{n}(\hat{\alpha} - \alpha)$ is asymptotically $\mathcal{N}(\mathbf{0}, \Sigma_{\text{d}}(\hat{\alpha}))$; that is $\sqrt{n}(\hat{\alpha} - \alpha) \sim \mathcal{AN}(\mathbf{0}, \Sigma_{\text{d}}(\hat{\alpha}))$, say, for certain $\Sigma_{\text{d}}(\hat{\alpha})$, which is a function of α for the estimator $\hat{\alpha}$.

It is natural to estimate α by such discriminative learning; however, the estimation is hindered by computational complexity related to $\sum_{i=1}^n \log(1 + e^{g(\mathbf{x}_i,\alpha)})$. Traditionally, generative learning is more commonly used.

1.1.4 Generative Learning

Generative learning uses the likelihood $\mathcal{L}$ and log-likelihood ℓ based on $p(\mathbf{x}, y)$, which are, respectively,

$$\mathcal{L}_{\mathrm{g}}(\theta) = \prod_{i=1}^{n} p(\mathbf{x}_i, y_i) = \prod_{i=1}^{n} p(y_i)p(\mathbf{x}_i|y_i) = \prod_{i=1}^{n} (\pi_1 p(\mathbf{x}_i|\mathcal{C}_1))^{y_i} (\pi_0 p(\mathbf{x}_i|\mathcal{C}_0))^{1-y_i} ,$$

$$\ell_{\mathrm{g}}(\theta) = \log \mathcal{L}_{\mathrm{g}}(\theta) = \sum_{i=1}^{n} y_i \log(\pi_1 p(\mathbf{x}_i|\theta_1)) + \sum_{i=1}^{n} (1 - y_i) \log(\pi_0 p(\mathbf{x}_i|\theta_0)) ,$$

where $p(\mathbf{x}_i|\theta_1) = p(\mathbf{x}_i|\mathcal{C}_1)$, $p(\mathbf{x}_i|\theta_0) = p(\mathbf{x}_i|\mathcal{C}_0)$, θ_1 and θ_0 are parameters of $p(\mathbf{x}|\mathcal{C}_1)$ and $p(\mathbf{x}|\mathcal{C}_0)$ for the two sub-populations $\mathcal{C}_1$ and $\mathcal{C}_0$, respectively, and θ is the vector of distinct elements within $\{\pi_1, \theta_1, \theta_0\}$.

Similarly, maximization of $\ell_{\mathrm{g}}(\theta)$, with respect to θ, leads to an estimator $\hat{\theta}$ of θ with $\sqrt{n}(\hat{\theta} - \theta) \sim \mathcal{AN}(\mathbf{0}, \Sigma_{\mathrm{g}}(\hat{\theta}))$, for certain $\Sigma_{\mathrm{g}}(\hat{\theta})$. However, we need to derive a generative estimator $\hat{\alpha}$ of α with $\sqrt{n}(\hat{\alpha} - \alpha) \sim \mathcal{AN}(\mathbf{0}, \Sigma_{\mathrm{g}}(\hat{\alpha}))$. The covariance matrix $\Sigma_{\mathrm{g}}(\hat{\theta})$ (or $\Sigma_{\mathrm{g}}(\hat{\alpha})$) is a function of θ (or α) for the estimator $\hat{\theta}$ (or $\hat{\alpha}$).

By Bayes' Theorem, $g(\mathbf{x}, \alpha) = \log \frac{p(y=1|\mathbf{x})}{p(y=0|\mathbf{x})} = \log \frac{\pi_1 p(\mathbf{x}|\theta_1)}{\pi_0 p(\mathbf{x}|\theta_0)}$, and thus the mapping $\alpha(\theta)$ and the relationship between $(\hat{\alpha} - \alpha)$ and $(\hat{\theta} - \theta)$ can be constructed. For example,

- if $\mathbf{x}|\theta_1 \sim \mathcal{N}(\mu_1, \Sigma)$, $\mathbf{x}|\theta_0 \sim \mathcal{N}(\mu_0, \Sigma)$, then

$$g(\mathbf{x}, \alpha) = \beta_0 + \beta^T \mathbf{x} = \log \frac{\pi_1}{\pi_0} - \frac{1}{2}(\mu_1^T \Sigma^{-1} \mu_1 - \mu_0^T \Sigma^{-1} \mu_0) + (\mu_1 - \mu_0)^T \Sigma^{-1} \mathbf{x} ; \quad (1.2)$$

- if $\mathbf{x}|\theta_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$, $\mathbf{x}|\theta_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$, then

$$g(\mathbf{x}, \alpha) = \beta_0 + \beta^T \mathbf{x} + \mathbf{x}^T \Gamma \mathbf{x} = \log \frac{\pi_1}{\pi_0} - \frac{1}{2}(\mu_1^T \Sigma_1^{-1} \mu_1 - \mu_0^T \Sigma_0^{-1} \mu_0) - \frac{1}{2} \log \frac{|\Sigma_1|}{|\Sigma_0|} +$$
$$(\mu_1^T \Sigma_1^{-1} - \mu_0^T \Sigma_0^{-1}) \mathbf{x} - \frac{1}{2} \mathbf{x}^T (\Sigma_1^{-1} - \Sigma_0^{-1}) \mathbf{x} . \quad (1.3)$$

The estimation of θ and thus α is hindered by potential mis-specification of sub-population densities.

1.2 Comparison between Generative and Discriminative Classifiers

For the generative classifiers, although maximum likelihood based on $p(\mathbf{x}, y; \theta)$ will lead to an asymptotically unbiased and efficient estimator $\hat{\theta}$ and consequently $\hat{\alpha}$, it can only be justified if $p(\mathbf{x}, y)$ is correctly specified. Similarly, for the discriminative classifiers, although

maximum likelihood based on $p(y|\mathbf{x};\alpha)$ will lead to an asymptotically unbiased and efficient estimator $\hat{\alpha}$, it can only be justified if $p(y|\mathbf{x})$ or, for example for the case of two classes y_1 and y_2, the corresponding discriminant function, $g(\mathbf{x},\alpha) = \log \frac{p(y_1|\mathbf{x})}{p(y_2|\mathbf{x})}$, is correctly specified. Different $p(\mathbf{x},y;\theta)$'s may lead to the same discriminant function $g(\mathbf{x},\alpha)$, which indicates that the discriminative classifiers may be less sensitive than the generative classifiers to the mis-specification of $p(\mathbf{x},y;\theta)$.

Comparison of generative and discriminative classifiers is an ever-lasting topic (Efron, 1975; O'Neill, 1980; Titterington et al., 1981; Rubinstein and Hastie, 1997; Ng and Jordan, 2001). In practice, commonly used discriminative and generative classifiers are logistic regression and normal-based discriminant analysis, respectively. Numerous theoretical, simulated and empirical comparisons between these two approaches have been investigated; see for example Efron (1975) and Titterington et al. (1981).

In general, the performance of such approaches depends on the correctness of the modelling, the bias, efficiency and consistency of the learning, and the reliability of the training data. For instance, when the modelling of $p(y|\pi)$ and $p(\mathbf{x}|y;\theta_{\mathrm{g}})$ is correct, normal-based linear discriminant analysis (LDA) can be more efficient than linear logistic regression (LLR) (Efron, 1975). However, the latter can perform better than the former when $\mathbf{x}|y$ is not normally distributed, because the latter does not necessarily assume the Gaussian form of $p(\mathbf{x}|y;\theta_{\mathrm{g}})$; for instance, the modelling of the latter is valid under general exponential family assumptions on $p(\mathbf{x}|y;\theta_{\mathrm{g}})$ (Efron, 1975).

Ng and Jordan (2001) presented some theoretical and empirical comparisons between LLR and the naïve Bayes classifier, a generative approach equivalent to LDA, when statistically independent and normally distributed features $\mathbf{x}$ within classes y are assumed. Their results suggested that, between the two approaches, there were the 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. Chapter 2 of this thesis will provide extension of and make comments on their study.

1.3 Combination of Generative and Discriminative Classifiers

If we consider the pros and cons of both discriminative and generative approaches (Efron, 1975; Titterington et al., 1981; Rubinstein and Hastie, 1997; Ng and Jordan, 2001), it is natural to exploit the best of both worlds. In this direction, many interesting proposals of hybrid learning techniques have emerged for combining the generative and discriminative approaches, such as the mixed discriminants (Rubinstein, 1998), the hybrid generative-discriminative algorithm (Raina et al., 2003; Fujino et al., 2007), the mixed log-likelihood (or the generative-discriminative tradeoff) (Rubinstein, 1998; Bouchard and Triggs, 2004), multi-conditional learning (McCallum et al., 2006) and a Bayesian blending (Bishop and Lasserre, 2007). Since the generative approaches can model unlabelled observations $\mathbf{x}_{1:m} = \{\mathbf{x}_j\}_{j=1}^m$ while the discriminative approaches do not, some of the above generative-discriminative combinations have been applied to semi-supervised learning scenarios (Suzuki et al., 2007; Druck et al., 2007; Bishop and Lasserre, 2007; Bouchard, 2007).

1.3.1 Hybrid Learning

Rubinstein (1998) presented the method of mixed discriminants, which involved constructing a discriminant $\hat{p}(y|\mathbf{x})$ by combining two posterior probabilities $p(y|\mathbf{x})$ obtained from a generative approach and a discriminative approach, respectively, as

$$\hat{p}(y=1|\mathbf{x}) = \lambda \frac{\exp(g(\mathbf{x}, \hat{\alpha}_{\mathrm{g}}))}{1+\exp(g(\mathbf{x}, \hat{\alpha}_{\mathrm{g}}))} + (1-\lambda)\frac{\exp(g(\mathbf{x}, \hat{\alpha}_{\mathrm{d}}))}{1+\exp(g(\mathbf{x}, \hat{\alpha}_{\mathrm{d}}))} ,$$

where $\lambda \in [0,1]$, and $\hat{\alpha}_{\mathrm{g}}$ and $\hat{\alpha}_{\mathrm{d}}$ are the generative and discriminative estimators of α, respectively. Since $\hat{\alpha}_{\mathrm{g}}$ and $\hat{\alpha}_{\mathrm{d}}$ can be estimated separately, this procedure is by nature similar to the construction of a new likelihood $\mathcal{L}_\lambda(\alpha, \theta)$ as a linear combination of two likelihoods $\mathcal{L}_{\mathrm{d}}$ and $\mathcal{L}_{\mathrm{g}}$ as $\mathcal{L}_\lambda(\alpha, \theta) = \lambda\mathcal{L}_{\mathrm{g}}(\theta) + (1-\lambda)\mathcal{L}_{\mathrm{d}}(\alpha)$, which may make the relationship between α and θ fail to comply with Bayes' Theorem.

McCallum et al. (2006) introduced the multi-conditional learning framework, one case of which defined a new log-likelihood $\ell_{\mathrm{MC}}(\theta) = \lambda_1 \ell_{\mathbf{x}|y}(\theta) + \lambda_2 \ell_{y|\mathbf{x}}(\theta)$, where $\ell_{\mathbf{x}|y}(\theta)$ and $\ell_{y|\mathbf{x}}(\theta)$ are log-likelihoods based on $p(\mathbf{x}|y)$ and $p(y|\mathbf{x})$, respectively, as functions of a common parameter vector θ. As pointed out by McCallum et al. (2006), this model is sensitive to the values of λ_1 and λ_2. With both $p(\mathbf{x}|y;\theta)$ and $p(y|\mathbf{x};\theta)$ derived from the joint distribution $p(\mathbf{x}, y; \theta)$, this model is a generative model with hybrid learning of θ.

Bishop and Lasserre (2007) provided a constructive Bayesian perspective to accommodate both the generative learning and discriminative learning of a generative model. This perspective adopted two parameter vectors θ_{d} and θ_{m} to describe the likelihood $\mathcal{L}_{\mathrm{g}}(\theta_{\mathrm{d}}, \theta_{\mathrm{m}})$ based on the joint distribution $p(\mathbf{x}, y; \theta_{\mathrm{d}}, \theta_{\mathrm{m}})$:

$$\mathcal{L}_{\mathrm{g}}(\theta_{\mathrm{d}}, \theta_{\mathrm{m}}) = p(\theta_{\mathrm{d}}, \theta_{\mathrm{m}}) \mathcal{L}_{y|\mathbf{x}}(\theta_{\mathrm{d}}) \mathcal{L}_{\mathbf{x}}(\theta_{\mathrm{m}}) \ ,$$

where θ_{d} and θ_{m} are parameters of a conditional distribution $p(y|\mathbf{x})$ and a mixture $p(\mathbf{x})$, respectively. This model implies that $p(y|\mathbf{x}; \theta_{\mathrm{d}}, \theta_{\mathrm{m}}) = p(y|\mathbf{x}; \theta_{\mathrm{d}})$ and $p(\mathbf{x}; \theta_{\mathrm{d}}, \theta_{\mathrm{m}}) = p(\mathbf{x}; \theta_{\mathrm{m}})$. Meanwhile, both $p(y|\mathbf{x}; \theta_{\mathrm{d}})$ and $p(\mathbf{x}; \theta_{\mathrm{m}})$ are derived from $p(\mathbf{x}, y; \theta_{\mathrm{d}}, \theta_{\mathrm{m}})$. When $\theta_d = \theta_m = \theta$ and the prior $p(\theta_{\mathrm{d}}, \theta_{\mathrm{m}})$ is uniform, this model corresponds to classical generative learning with $\mathcal{L}_{\mathrm{g}}(\theta)$.

As a result of its representation of $p(\mathbf{x}, y)$ in terms of $p(y|\mathbf{x})$ and $p(\mathbf{x})$, such a Bayesian blending can be naturally employed for semi-supervised learning, where the labelled observations are used for $\mathcal{L}_{y|\mathbf{x}}(\theta_{\mathrm{d}})$ and the unlabelled observations for $\mathcal{L}_{\mathbf{x}}(\theta_{\mathrm{m}})$ (Bishop and Lasserre, 2007). For semi-supervised learning but derived from the multi-conditional learning framework, Druck et al. (2007) proposed a related model, which uses a common θ to define a new log-likelihood $\ell_{\mathrm{MC}^*}(\theta) = \lambda_1 \ell_{\mathbf{x}}(\theta) + \lambda_2 \ell_{y|\mathbf{x}}(\theta)$. Considering the non-convexity of $\ell_{\mathbf{x}}(\theta)$ and the difference between the scales of $\ell_{\mathbf{x}}(\theta)$ and $\ell_{y|\mathbf{x}}(\theta)$, the model is also sensitive to the determination of λ_1 and λ_2. In addition, Druck et al. (2007) provided empirical comparison between their model and that of Bishop and Lasserre (2007).

Raina et al. (2003) and Fujino et al. (2007) proposed the hybrid generative-discriminative algorithm, which partitions the feature vector $\mathbf{x}$ into multiple partial vectors with different weights θ_{d}. This leads to a parameter vector $(\theta^T, \theta_{\mathrm{d}}^T)^T$, where θ is estimated generatively while θ_{d} is estimated discriminatively. It can be regarded as a generative model using both generative and discriminative learning, in the sense that it assumes a scaled $p(\mathbf{x}|y; \theta, \theta_{\mathrm{d}})$ and the discriminative learning of θ_{d} is based on the estimation of θ (see Chapter 4 for details).

The focus of Chapter 3 of this thesis is on an alternative hybrid learning method, the generative-discriminative tradeoff approach (GDT, or the mixed log-likelihood method) (Rubinstein, 1998; Bouchard and Triggs, 2004). The GDT constructs a new log-likelihood as a weighted average of the log-likelihoods $\ell_{\mathrm{g}}(\theta)$ for generative learning and $\ell_{\mathrm{d}}(\alpha)$ for discriminative learning, given by $\ell_\lambda(\theta, \alpha) = \lambda \ell_{\mathrm{g}}(\theta) + (1 - \lambda) \ell_{\mathrm{d}}(\alpha)$, for $0 < \lambda < 1$. In order to couple the two separate estimations of $\hat{\theta}$ and $\hat{\alpha}$, either θ should be rewritten as a function $\theta(\alpha)$ of α, or

α as a function $\alpha(\theta)$ of θ. In general, $p(y|\mathbf{x})$ can be derived from $p(\mathbf{x}, y)$, but not vice versa, and the dimension of θ is larger than that of α, as with LDA. Therefore, it is more feasible to use $\alpha(\theta)$ and thus only the parameter vector θ remains in the new log-likelihood:

$$\ell_\lambda(\theta) = \lambda \ell_{\mathrm{g}}(\theta) + (1 - \lambda)\ell_{y|\mathbf{x}}(\theta) \ ,$$

where, as defined earlier, $\ell_{\mathrm{g}}(\theta) = \sum_{i=1}^{n} \log p(\mathbf{x}_i, y_i)$, while

$$\ell_{y|\mathbf{x}}(\theta) = \sum_{i=1}^{n} \log p(y_i|\mathbf{x}_i) = \sum_{i=1}^{n} \log \frac{\pi_{y_i} p(\mathbf{x}_i|y_i; \theta_{y_i})}{\pi_1 p(\mathbf{x}_i|\theta_1) + \pi_0 p(\mathbf{x}_i|\theta_0)} \ ,$$

a discriminative log-likelihood, but as a function of θ rather than α.

As with other hybrid learning techniques, the GDT is modelled through $p(y|\pi)$ and $p(\mathbf{x}|y; \theta_{\mathrm{g}})$ and thus is by nature a generative model with hybrid learning, learning the common θ within both likelihoods. The GDT has, through combination with the hybrid generative-discriminative algorithm (Raina et al., 2003), also been used for semi-supervised learning (Suzuki et al., 2007).

All these hybrid learning techniques demonstrated in practice that their classification performance could be superior to the generative component or the discriminative component alone.

1.4 Generative and Discriminative Hidden Markov Models

Amongst the latent (hidden) variable models for structured data such as time series, hidden Markov models (HMMs) for discrete-valued hidden states and state-space models (SSMs) for continuous-valued hidden states are widely used.

Traditionally, an HMM is generative because it models a distribution $P(O_1^n|S_1^n)$, the DGP of the observed output sequence, $O_1^n = o_1, \ldots, o_n$, given the hidden state sequence, $S_1^n = s_1, \ldots, s_n$, and thus $P(O_1^n|S_1^n)$, a state-dependent term, is included in the criterion for determining a stochastic optimal sequence of hidden states. Recently, Zhou (2005) proposed a discriminative hidden Markov model (D-HMM), which includes output-dependent terms $P(s_t|O_1^n), t = 1, \ldots, n$, in the criterion, based on an assumption of "mutual information independence". Meanwhile, Li (2005) presented the so-called "hidden Markov models with states depending on observations" (HMMSDO), which assume that the current state s_t depends not only on the last state s_{t-1} but also on the last output o_{t-1}, so that output-dependent terms $P(s_t|s_{t-1}, o_{t-1})$ are included in the criterion.

Both the D-HMM and HMMSDO show superior performance in determining the optimal state sequence for certain applications. Zhou (2005) shows that the D-HMM outperforms

the corresponding generative hidden Markov model (G-HMM) for part-of-speech tagging and phrase chunking; Li (2005) shows that HMMSDO outperforms the standard HMM for prediction of protein secondary structures when the training set is large enough.

Chapter 6 will study the assumption of "mutual information independence" and will extend it to derive generative (state-dependent) representations of these two discriminative (output-dependent) HMMs.

1.5 Generative Approaches to Image Thresholding

Image thresholding is a simple and widely-used technique for segmentation, partitioning a grey-level image into segments corresponding to different classes (Sahoo et al., 1988; Pal and Pal, 1993; Sezgin and Sankur, 2004), given that the classes to some extent can be distinguished by their grey levels. Most thresholding approaches are proposed for two-class binarisation and are based on the grey-level histogram of an image (Sahoo et al., 1988; Sezgin and Sankur, 2004; Glasbey, 1993; Trier and Jain, 1995). Two of the most popular approaches are Otsu's method (Otsu, 1979) and Kittler and Illingworth's minimum error thresholding (MET) (Kittler and Illingworth, 1986).

Kurita et al. (1992) show that Otsu's method is equivalent to maximisation of the log-likelihood based on the conditional distribution $p(x|y)$, where x is the grey level and $y \in \{0, 1\}$ is the class indicator corresponding to x, under the assumption that the grey level within each class (denoted by $x|y$) follows a normal distributions $\mathcal{N}(\mu_y, \sigma_y^2)$ and $\sigma_0^2 = \sigma_1^2$. Kurita et al. (1992) also shows that MET is equivalent to maximisation of the log-likelihood based on the joint distribution $p(x, y)$, under the assumption that $x|y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ and $\sigma_0^2 \neq \sigma_1^2$. Since $p(x, y) = \pi_y p(x|y)$, where $\pi_y = p(y)$, Otsu's method is also equivalent to maximisation of the log-likelihood based on $p(x, y)$ with $\pi_0 = \pi_1 = 0.5$. In this sense, both Otsu's method and MET assume a DGP $p(x, y)$; therefore, we call such approaches generative thresholding approaches. As with Fisher's linear discriminant, the Otsu's original method does not assume normally distributed classes or that $\sigma_0^2 = \sigma_1^2$; therefore, hereafter we refer, as Otsu's method, to the generative method to which it is equivalent, shown in Kurita et al. (1992). In Chapter 7, we will propose discriminative extensions of the traditional generative approaches to thresholding.

1.6 Contributions of this Thesis

The contributions of this thesis are the following.

First, Ng and Jordan (2001) claimed that there exist two distinct regimes of performance between the generative and discriminative classifiers with regard to the training-set size; however, our empirical and simulation studies, as presented in Chapter 2, suggest that it is not so reliable to claim such an existence of the two distinct regimes. 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 you have in the correctness of the specification of either $p(y|\mathbf{x})$ or $p(\mathbf{x}, y)$. This can be to some extent a demonstration of why Efron (1975) and O'Neill (1980) prefer LDA but other empirical studies may prefer LLR instead. Furthermore, we suggest that pairing of either LDA assuming a common diagonal covariance matrix (LDA-Λ) or the naïve Bayes classifier and LLR 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 LLR to be generalised to all generative and discriminative classifiers.

Secondly, in Chapter 3, we present the interpretation and asymptotic relative efficiency (ARE) of the GDT approach for linear and quadratic normal discrimination without model mis-specification, and compare its ARE with those of its generative and discriminative counterparts. The classification performance of the GDT is compared with those of LDA and LLR on simulated datasets. We argue that the GDT is a generative model integrating both discriminative and generative learning. It is therefore sensitive to model mis-specification of the data-generating process and, in practice, its discriminative component may behave differently from a truly discriminative approach. Amongst the three approaches that we compare, the asymptotic efficiency of the GDT is lower than that of the generative approach when no model mis-specification occurs. In addition, without model mis-specification, LDA performs the best; with model mis-specification, the GDT may perform the best at an optimal tradeoff between its discriminative and generative components, and LLR, a truly discriminative classifier, in general performs well when the training-sample size is reasonably large.

Thirdly, in Chapter 4, we interpret the hybrid algorithm from three perspectives, namely class-conditional probabilities, class-posterior probabilities and loss functions underlying the model. We suggest that the hybrid algorithm is by nature a generative model with its parame-

ters learnt through both generative and discriminative approaches, in the sense that it assumes a scaled data-generation process and uses scaled class-posterior probabilities to perform discrimination. Our suggestion can also be applied to its multi-class extension. In addition, using simulated and real-world data, we compare the performance of the normalised hybrid algorithm as a classifier with that of the naïve Bayes classifier and LLR. Our simulation studies suggest in general the following: if the covariance matrices are diagonal matrices, the naïve Bayes classifier performs the best; if the covariance matrices are full matrices, LLR performs the best. Our studies also suggest that the hybrid algorithm may provide worse performance than either the naïve Bayes classifier or LLR alone.

Fourthly, based on our studies presented in Chapters 2, 3 and 4, we propose in Chapter 5 a joint generative-discriminative modelling (JGD) approach to classification, by partitioning variables into two subsets based on statistical tests of the DGP. Our JGD approach adopts statistical tests, such as normality tests, of the assumed DGP for each variable to justify the use of generative approaches for the variables which satisfy the tests and of discriminative approaches for other variables. Such a partition of variables and a combination of generative and discriminative approaches are derived in a probabilistic rather than a heuristic way. We have concentrated on particular choices for the generative and discriminative components of our models, but the overall principle is quite general and can accommodate many other special versions. Of course, we must ensure that the assumptions underlying the resulting generative classifiers can be tested statistically. Numerical results from real UCI and gene-expression data and from simulated data demonstrate promising performance of this new approach for practical application to both low- and high-dimensional data.

Fifthly, in Chapter 6, we study the assumption of "mutual information independence", which is used by Zhou (2005) for deriving the so-called discriminative HMM (D-HMM). We suggest that the mutual information assumption (6.6) results in the D-HMM, while another mutual information assumption (6.12) results in its generative counterpart, the G-HMM. However, in practice, whether or not the assumptions are reasonable and how the corresponding HMMs perform can be data-dependent; research efforts to explore an adaptive switching between or combination of these two models may be worthwhile. Meanwhile, we suggest that the so-called output-dependent HMMs could be represented in a state-dependent manner, and vice versa, essentially by application of Bayes' theorem.

Finally, in Chapter 7, we present discriminative approaches to histogram-based image

thresholding, in which the optimal threshold is derived from the maximum likelihood based on the conditional distribution $p(y|x)$ of y, the class indicator of a grey level x, given x. The discriminative approaches can be regarded as discriminative extensions of the traditional generative approaches to thresholding, such as Otsu's method and Kittler and Illingworth's MET. As illustrations, we develop discriminative versions of Otsu's method and MET by using discriminant functions corresponding to the original methods to represent $p(y|x)$. These two discriminative thresholding approaches are compared with their original counterparts on selecting thresholds for a variety of histograms of mixture distributions. Results show that the discriminative Otsu method consistently provides relatively good performance. Although being of higher computational complexity than the original methods in parameter estimation, its robustness and model simplicity can justify the discriminative Otsu method for scenarios in which the risk of model mis-specification is high and the computation is not demanding.

Chapter 2

Comparison between Generative and Discriminative Models

In this chapter, we first replicate and extend experiments on the 15 real-world datasets used by Ng and Jordan (2001), for empirical comparison between LDA-Λ or the naïve Bayes classifiers and linear logistic regression (LLR). Then, as Ng and Jordan (2001) claim that there are two distinct regimes of performance with regard to the training-set size, we clarify such a claim further through commenting on the reliability of the two regimes and the parity between the compared classifiers.

2.1 Introduction

Comparison of generative and discriminative classifiers is an ever-lasting topic (Efron, 1975; O'Neill, 1980; Titterington et al., 1981; Rubinstein and Hastie, 1997; Ng and Jordan, 2001).

Ng and Jordan (2001) presented some theoretical and empirical comparisons between linear logistic regression and the naïve Bayes classifier. The naïve Bayes classifier is a generative classifier, which assumes statistically independent features $\mathbf{x}$ within classes y and thus diagonal covariance matrices within classes; it is equivalent to normal-based linear (for a common diagonal covariance matrix) or quadratic (for unequal diagonal within-class covariance matrices) discriminant analysis, when $\mathbf{x}$ is assumed normally distributed for each class. The results in Ng and Jordan (2001) 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.

The setting for the theoretical proof and empirical evidence in Ng and Jordan (2001) 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.

In the case of discrete features, each feature $x_i, i = 1, \ldots, p$, independent of other features within $\mathbf{x}$, is assumed within a class to be a binomial variable such that its value $x_i \in \{0, 1\}$ within each class. We observe, however, this may not guarantee the discriminant function $\lambda(\alpha) = \log\{p(y = 1|\mathbf{x})/p(y = 2|\mathbf{x})\}$, where α is a parameter vector, to be linear; therefore, the naïve Bayes classifier may not be a partner of linear logistic regression as a generative-discriminative pair.

In the case of continuous features, $\mathbf{x}|y$ 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. All of the observed values of the features are rescaled so that $x_i \in [0, 1]$.

Based on such a setting, Ng and Jordan (2001) compared two so-called generative-discriminative pairs: one is for the continuous case, comparing LDA assuming a common diagonal covariance matrix Λ (denoted by LDA-Λ hereafter) vs. linear logistic regression, and the other is for the discrete case, comparing the naïve Bayes classifier vs. linear logistic regression.

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 (denoted by LDA-Σ hereafter) vs. linear logistic regression also merits investigation. In addition, 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 Ng and Jordan (2001).

Ng and Jordan (2001) 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 (Asuncion and Newman, 2007); 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 Ng and Jordan (2001), 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 large enough training sets. However, it is known that the performance of a classifier varies to some extent with the features selected.

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 (Verboven and Hubert, 2005). 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 Ng and Jordan (2001), 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 sample for each of the two classes, and, for discrete datasets, to have all the levels of the features presented by the training samples, 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.

Meanwhile, we observe that, 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 samples 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.

2.2 Linear Discrimination On Continuous Datasets

For the continuous datasets, as was done by Ng and Jordan (2001), 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, before carrying out the classification, we perform the Shapiro-Wilk test for within-class normality for each feature $x_i|y$ and Levene's test for homogeneity of variance across the two classes. Levene's test is less sensitive to deviations from normality than is the Bartlett test, another test for homogeneity of variance. For the following datasets, 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.

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

Table 2.1: Description of continuous datasets.

A brief description of the continuous datasets can be found in Table 2.1, 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 fea-

tures, 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 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 2.4), 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 2.1. Since with variable selection by AIC the results conform more to the claim of two regimes by Ng and Jordan (2001), we show such results if they are different from those without variable selection. Meanwhile, in the figures hereafter we use the same annotations of the vertical and horizontal axes and the same line type as those in Ng and Jordan (2001). All the observations from these figures are only valid for $m > p$, with the intercept in $\lambda(\alpha)$ taken into account.

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

1. In the comparison of LDA-Λ vs. linear logistic regression, the pattern of our results can be said to be similar to that of Ng and Jordan (2001).

2. 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.

3. 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.

4. Pre-processing with variable selection can reveal the distinction in performance of generative and discriminative classifiers with fewer training samples.

5. 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.

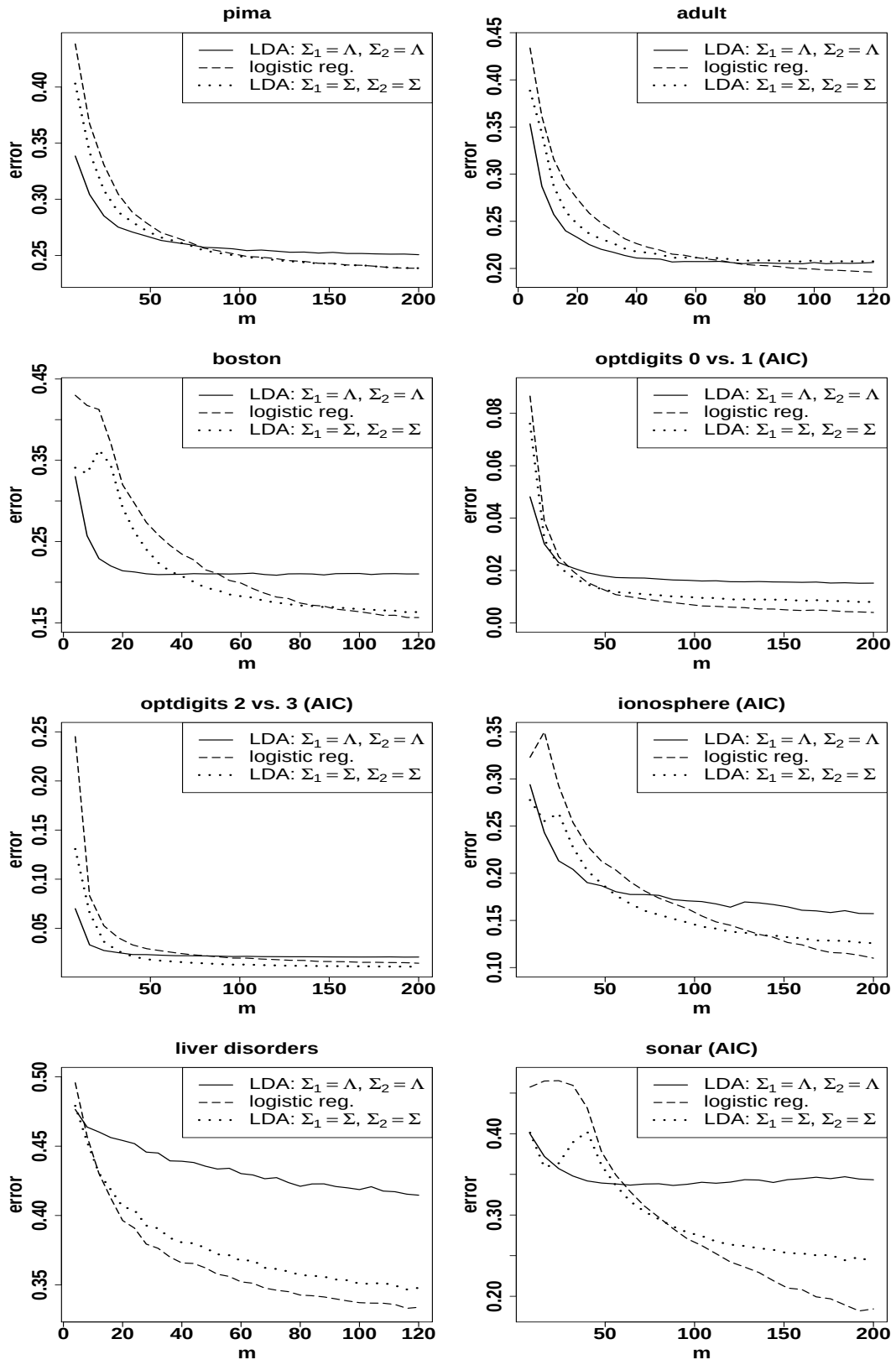


Figure 2.1: Plots of misclassification error rate vs. training-set size m (averaged over 1000 random training/test set splits) on the continuous UCI datasets, with regard to linear discrimination.

2.3 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 2.2, 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 2.2 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 2.1, 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.

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 2.2, suggests quite similar conclusions to those in Section 2.3, through substituting QDA-Λ_g for LDA-Λ, QDA-Σ_g for LDA-Σ, and quadratic logistic regression for linear logistic regression.

2.4 Linear Discrimination On Discrete Datasets

For the discrete datasets, as was done by Ng and Jordan (2001), all the continuous features are removed and only the discrete features are used. The results are entitled ‘multinomial’ in following figures if a dataset includes multinomial features, and otherwise are entitled ‘binomial’. Meanwhile, any observation with missing features is removed from the datasets, as is any feature with only a single value for all the observations.

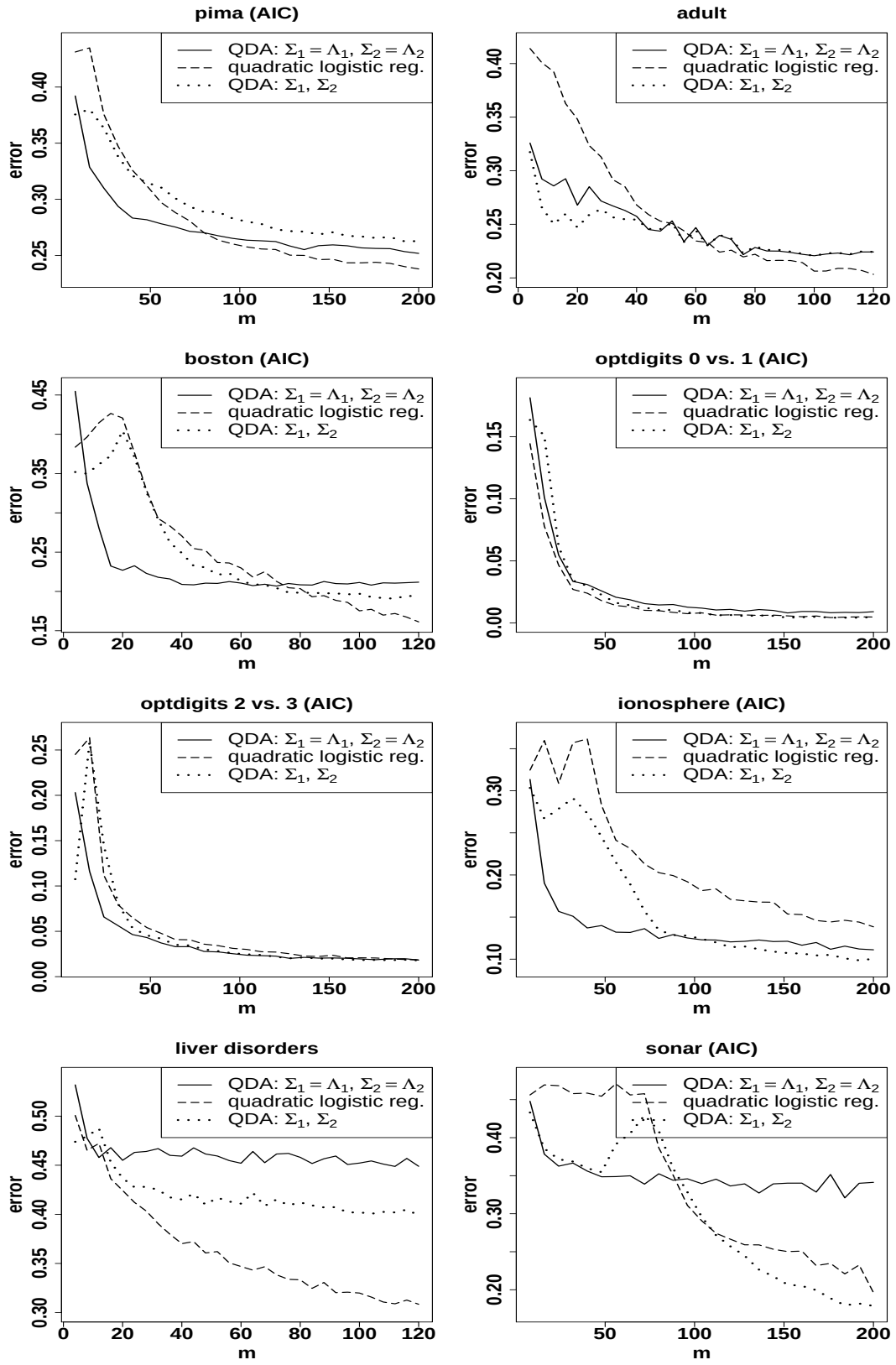


Figure 2.2: Plots of misclassification error rate vs. training-set size m (averaged over 100 random training/test set splits) on the continuous UCI datasets, with regard to quadratic discrimination.

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

Table 2.2: Description of discrete datasets.

A brief description of the discrete datasets can be found in Table 2.2, 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 2.3. All the observations from these figures are only valid for $m > \tilde{p}$, 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 Ng and Jordan (2001).

2.5 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 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

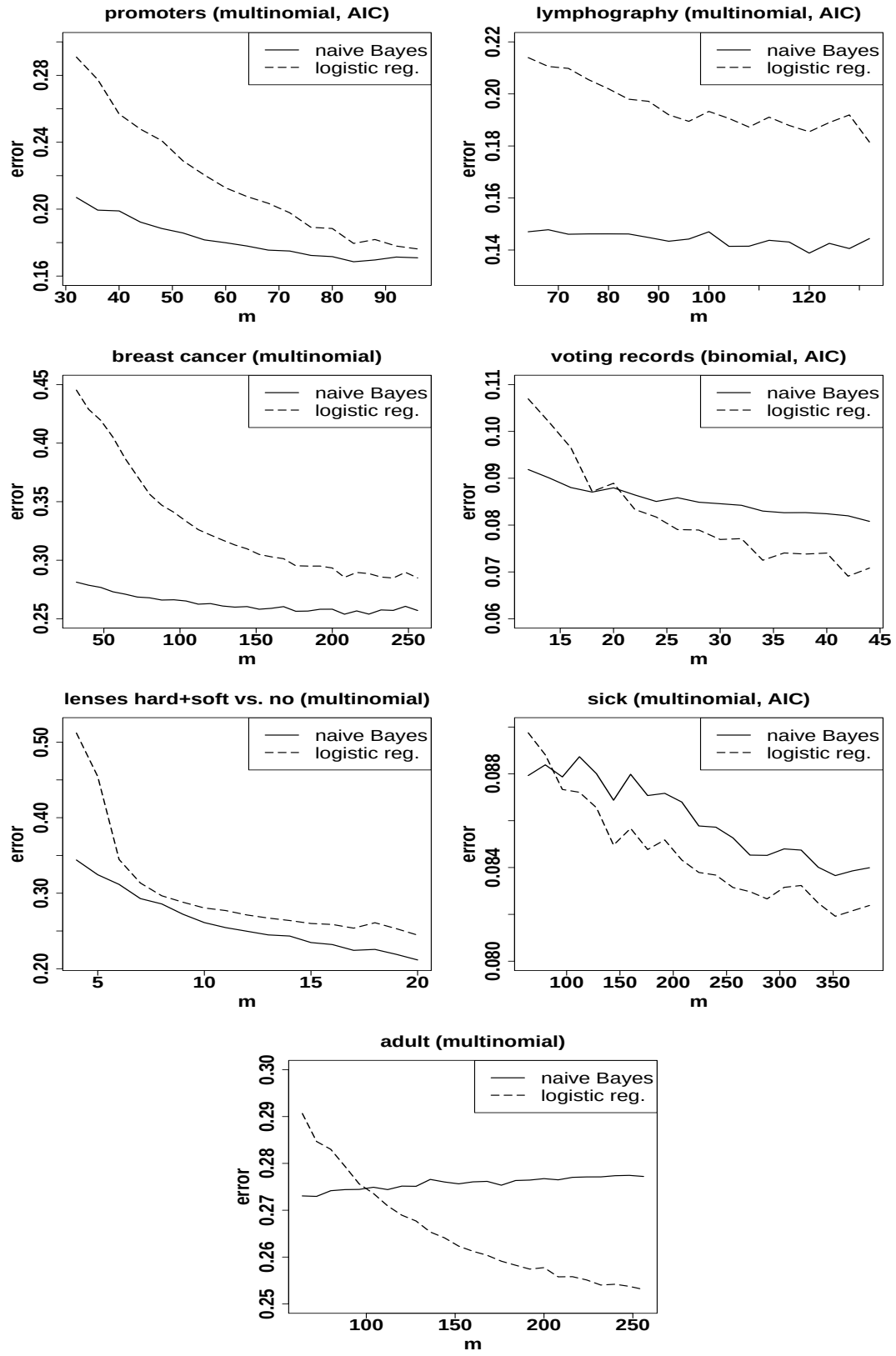


Figure 2.3: Plots of misclassification error rate vs. training-set size m (averaged over 1000 random training/test set splits) on the discrete UCI datasets, with regard to linear discrimination.

used since the feature space is only of dimension two.

2.5.1 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 2.4.

The dataset for the top-left panel of Figure 2.4 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.

2.5.2 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 2.5.1.

The results are shown in Figure 2.5, where for each panel the constraint with regard to Σ_1 and Σ_2 is the same as the corresponding one in Figure 2.4, except for a scalar multiplier

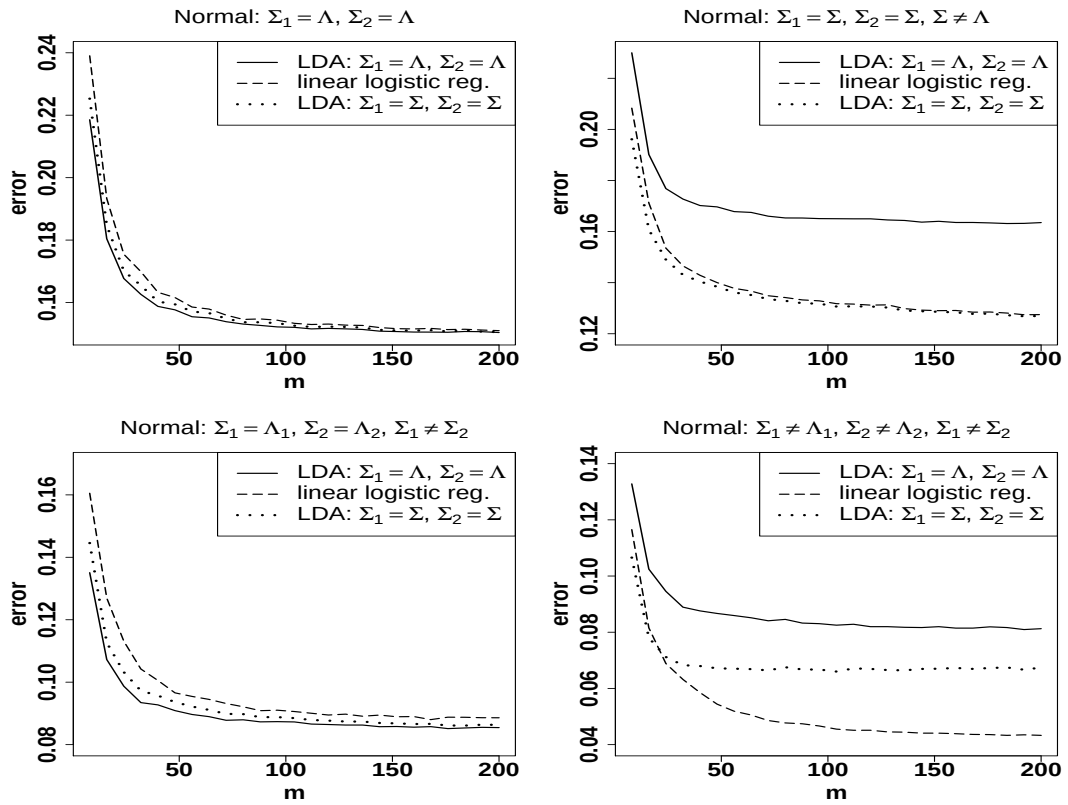


Figure 2.4: Plots of misclassification error rate vs. training-set size m (averaged over 1000 random training/test set splits) on simulated bivariate normally distributed data for two classes.

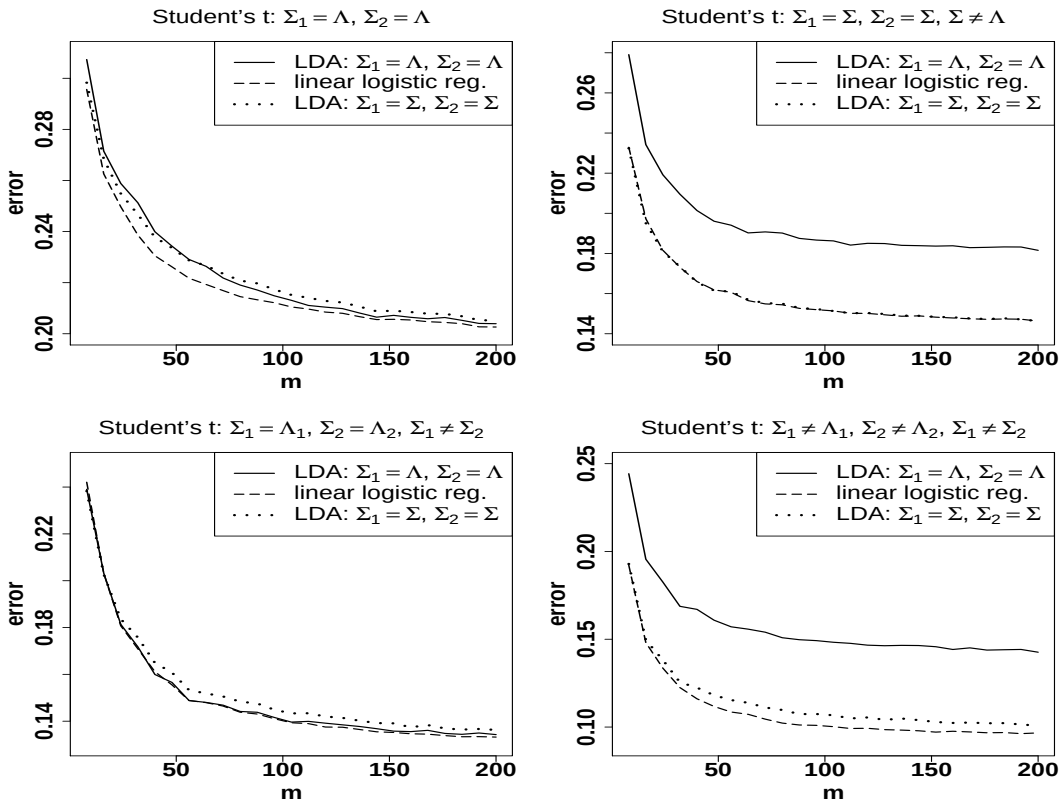


Figure 2.5: Plots of misclassification error rate vs. training-set size m (averaged over 1000 random training/test set splits) on simulated bivariate Student's t-distributed data for two classes.

$\nu/(\nu-2)$.

2.5.3 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 2.5.1.

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 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, considering 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.

The results are shown in Figure 2.6, where for each panel the constraint with regard to Σ_1 and Σ_2 is the same as the corresponding one in Figure 2.4.

2.5.4 Normal Mixture Data

Compared with the normal distribution, the Student's *t*-distribution and the log-normal distribution used in Sections 2.5.1, 2.5.2 and 2.5.3 for the comparison of the classifiers, the mixture

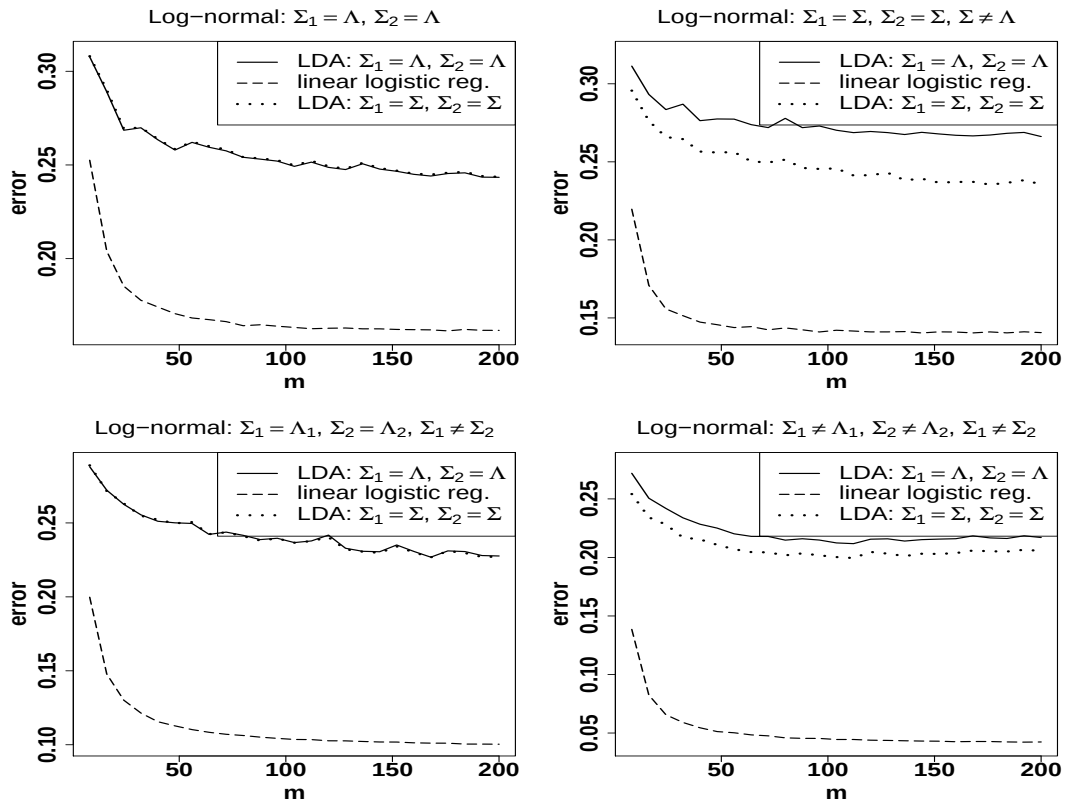


Figure 2.6: Plots of misclassification error rate vs. training-set size m (averaged over 1000 random training/test set splits) on simulated bivariate log-normally distributed data for two classes.

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 2.5.1.

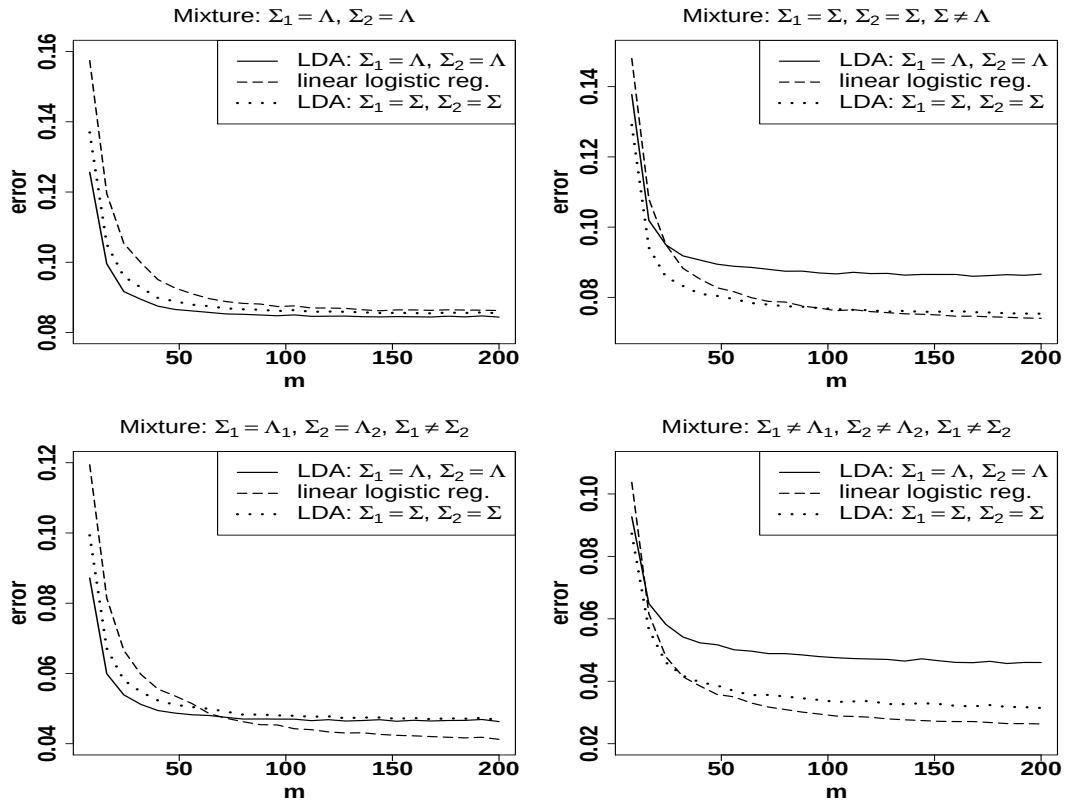


Figure 2.7: Plots of misclassification error rate vs. training-set size m (averaged over 1000 random training/test set splits) on simulated bivariate 2-component normal mixture data for two classes.

The results are shown in Figure 2.7, where for each panel the constraint with regard to Σ_1 and Σ_2 is the same as the corresponding one in Figure 2.4.

2.5.5 Summary of Linear Discrimination on Simulated Datasets

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

1. When the data are consistent with the assumptions underlying LDA-Λ or LDA-Σ, 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.

2. When the data violate the assumptions underlying the LDAs, linear logistic regression generally performs better than the LDAs, in particular when m is large; 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.

3. 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.

2.6 Comments on Comparison of Discriminative and Generative Classifiers

Based on the theoretical analysis and empirical comparison between LDA-Λ or the naïve Bayes classifiers and linear logistic regression, Ng and Jordan (2001) claim that there are two distinct regimes of performance with regard to the training-set size. Such a claim can be clarified further through commenting on the reliability of the two regimes and the parity between the compared classifiers.

2.6.1 On the Two Regimes of Performance regarding Training-Set Size

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 the 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 $\hat{y}(\mathbf{x}) = \operatorname{argmax}_y p(y|\mathbf{x})$, is used for the 2-class classification.

Discriminative classifiers estimate the parameter α of the discriminant function $\lambda(\alpha)$ through maximising a conditional probability $\operatorname{argmax}_\alpha p(\underline{y}_{tr}|\underline{\mathbf{x}}_{tr}, \alpha)$; 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. 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 2.5, can be proposed.

- 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.

- 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.

- We note that all of the above claims are based on the premise that 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.

- 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.

Generative classifiers estimate the parameter α of the discriminant function $\lambda(\alpha)$ through first maximising a joint probability $\text{argmax}_\theta\, p(\underline{y}_{tr}, \underline{\mathbf{x}}_{tr}|\theta)$ 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 2.5, can be proposed.

- Asymptotically, the generative classifiers will provide the best prediction performance on $(\underline{y}_{te}, \underline{\mathbf{x}}_{te})$. However, this is dependent on the premise that $p(y, \mathbf{x}|\theta)$ is correctly specified for all the observations.

- 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.

- We note that all of the above claims are based on the premise that that $p(y, \mathbf{x}|\theta)$ is correctly specified for all the observations.

- 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 $(\underline{y}_{te}, \underline{\mathbf{x}}_{te})$.

In summary, it is not 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 2.2 and 2.4, 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 $p(y, \mathbf{x})$. This can be to some extent a demonstration of

why Efron (1975) and O'Neill (1980) prefer LDA but other empirical studies may prefer linear logistic regression instead.

2.6.2 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 formulation 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 naïve Bayes classifier and LDA-Λ with a diagonal covariance matrix 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 side, LDA-Λ and the naïve Bayes classifier.

In this section, we will illustrate such pairing of two generative-discriminative pairs: one is LDA-Λ vs. linear logistic regression (Ng and Jordan, 2001), and the other is the naïve Bayes classifier vs. generalised additive model (GAM) (Rubinstein and Hastie, 1997).

2.6.2.1 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

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

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 $\mathbf{x}$.

In 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

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

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-Σ. 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 O'Neill (1980), 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 $\lambda_{\text{gen}}(\alpha)$.

2.6.2.2 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 as a sum of p functions $f(x_i),i=1,\ldots,p$, of the p features x_i separately (Rubinstein and Hastie, 1997); that is

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

Meanwhile, besides the assumption of the distribution of $(\mathbf{x}|y)$, a fundamental assumption underlying the naïve Bayes classifier is 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

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

It is clear, as pointed out by Rubinstein and Hastie (1997), 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.

In summary, considering 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 the generative and discriminative classifiers.

Chapter 3

On the Generative-Discriminative Tradeoff Approach

In this chapter, we first briefly introduce the generative-discriminative tradeoff method (GDT) (Rubinstein, 1998; Bouchard and Triggs, 2004; Bouchard, 2007) and present its interpretation, then compare its asymptotic efficiency with those of its generative and discriminative counterparts for linear and quadratic normal discrimination when there is no model mis-specification, and finally compare the performance of the GDT, LDA and LLR methods for two-class discrimination using simulated datasets.

3.1 Introduction

The GDT constructs a new log-likelihood as a weighted average of the log-likelihoods $\ell_{\mathrm{g}}(\theta)$ for generative learning and $\ell_{\mathrm{d}}(\alpha)$ for discriminative learning, given by $\ell_\lambda(\theta, \alpha) = \lambda\ell_{\mathrm{g}}(\theta) + (1-\lambda)\ell_{\mathrm{d}}(\alpha)$, for $0 < \lambda < 1$. In order to couple the two separate estimations of $\hat{\theta}$ and $\hat{\alpha}$, either θ should be rewritten as a function $\theta(\alpha)$ of α, or α as a function $\alpha(\theta)$ of θ. In general, $p(y|\mathbf{x})$ can be derived from $p(\mathbf{x}, y)$, but not vice versa, and the dimension of θ is larger than that of α, as with LDA. Therefore, it is more feasible to use $\alpha(\theta)$ and thus only the parameter vector θ remains in the new log-likelihood:

$$\ell_\lambda(\theta) = \lambda\ell_{\mathrm{g}}(\theta) + (1-\lambda)\ell_{y|\mathbf{x}}(\theta) \ ,$$

where, as defined earlier, $\ell_{\mathrm{g}}(\theta) = \sum_{i=1}^{n} \log p(\mathbf{x}_i, y_i)$ and $y_i \in \{0, 1\}$, while

$$\ell_{y|\mathbf{x}}(\theta) = \sum_{i=1}^{n} \log p(y_i|\mathbf{x}_i) = \sum_{i=1}^{n} \log \frac{\pi_{y_i} p(\mathbf{x}_i|y_i; \theta_{y_i})}{\pi_1 p(\mathbf{x}_i|\theta_1) + \pi_0 p(\mathbf{x}_i|\theta_0)} ,$$

a discriminative log-likelihood, but as a function of θ rather than α.

As with other hybrid learning techniques, the GDT is modelled through $p(y|\pi)$ and $p(\mathbf{x}|y; \theta_{\mathrm{g}})$ and thus is by nature a generative model with hybrid learning, learning the common θ within both likelihoods.

From a probabilistic point of view, if there exists a distribution

$$p(\mathbf{x}, y; \theta, \lambda) = c(\lambda) p(\mathbf{x}, y; \theta)^{\lambda} p(y|\mathbf{x}; \theta)^{1-\lambda} ,$$

then

$$\underset{\theta}{\operatorname{argmax}}\, \ell_{\lambda}(\theta) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(\mathbf{x}_i, y_i; \theta, \lambda) .$$

To justify that the GDT can be derived from a well-defined model, Bouchard (2007) provides a joint distribution

$$Q(\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}; \theta, \lambda) = (1 - \varsigma(\lambda)) \prod_{i=1}^{n} p(y_i|\mathbf{x}_i; \theta) U(\mathbf{x}_i) + \varsigma(\lambda) \prod_{i=1}^{n} p(\mathbf{x}_i, y_i; \theta) ,$$

where $U(\mathbf{x}_i)$ is not necessarily equal to $p(\mathbf{x}_i)$, and $\varsigma(\lambda)$ is a function satisfying

$$\underset{\theta}{\operatorname{argmax}}\, \ell_{\lambda}(\theta) = \underset{\theta}{\operatorname{argmax}}\, Q(\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}; \theta, \lambda) .$$

Some algebra shows that $\ell_{y|\mathbf{x}}(\theta) = \ell_{\mathrm{g}}(\theta) - \ell_{\mathbf{x}}(\theta)$, where $\ell_{\mathbf{x}}(\theta) = \sum_{i=1}^{n} \log(\pi_1 p(\mathbf{x}_i|\theta_1) + \pi_0 p(\mathbf{x}_i|\theta_0))$ is the log-likelihood of a 2-component mixture. It follows that, first, $\ell_{\lambda}(\theta) = \ell_{\mathrm{g}}(\theta) + (\lambda - 1)\ell_{\mathbf{x}}(\theta)$, which indicates that the GDT can be viewed as regularised generative learning; secondly, $\ell_{\lambda}(\theta) = \ell_{y|\mathbf{x}}(\theta) + \lambda \ell_{\mathbf{x}}(\theta)$, which indicates that the GDT can also be viewed as regularised discriminative learning; both regularisation penalties are determined by mixture data (Rubinstein, 1998). Furthermore, with $p(y)$ known and $\lambda_2 = 1 - \lambda_1$, the multi-conditional learning framework (McCallum et al., 2006) can be equivalent to the GDT with regard to parameter estimation.

Maximization of $\ell_{\lambda}(\theta)$, with respect to θ, leads to an estimator $\hat{\theta}$ of θ with $\sqrt{n}(\hat{\theta} - \theta) \sim \mathcal{AN}(\mathbf{0}, \Sigma_{\lambda}(\hat{\theta}))$, say, for certain $\Sigma_{\lambda}(\hat{\theta})$. Based on this, as in the generative approaches, we can derive the estimator $\hat{\alpha}$ of α with $\sqrt{n}(\hat{\alpha} - \alpha) \sim \mathcal{AN}(\mathbf{0}, \Sigma_{\lambda}(\hat{\alpha}))$, for certain $\Sigma_{\lambda}(\hat{\alpha})$.

In addition, encouraging results from two simulation experiments in Bouchard and Triggs (2004), in which the GDT assumes for the sub-populations two normal distributions with a

common diagonal covariance matrix Λ, imply, in the sense of minimum logistic loss, the following conclusions.

1. Without mis-specification, the generative component alone, with $\lambda = 1$, which in fact corresponds to LDA with a common Λ (hereafter denoted by LDA-Λ), has the best performance while the discriminative component alone, with $\lambda = 0$, has the worst performance.

2. With mis-specification, the performance of the discriminative component alone, with $\lambda = 0$, improves as the training-set size n increases, starting from being worse than that of the generative component alone to being better.

3. With mis-specification, the GDT, with $0 < \lambda < 1$, has the best performance, for certain λ.

We make the following observations: implication (1) conforms to the results of Efron (1975) and O'Neill (1980) that a generative model (LDA) enjoys better asymptotic classification performance than its discriminative counterpart (LLR); implication (2) conforms to the results of Ng and Jordan (2001); while implication (3) conforms at an abstract level to those of other hybrid learning techniques. In this chapter, we provide some theoretical support for implication (1), from the perspective of asymptotic relative efficiency (ARE) in terms of misclassification error rate, for linear and quadratic normal discrimination. Bouchard (2005) provided some asymptotic results in terms of logistic loss; nevertheless, for classification, the error rate is of more practical use than the logistic loss.

3.2 Asymptotic Efficiency of GDT

3.2.1 Asymptotic Relative Efficiency (ARE)

Given no mis-specification of the two sub-population densities, namely $p(\mathbf{x}|\theta_1)$ and $p(\mathbf{x}|\theta_0)$, the optimal boundary for classification should be $g(\mathbf{x}, \alpha) = \log \frac{\pi_1 p(\mathbf{x}|\theta_1)}{\pi_0 p(\mathbf{x}|\theta_0)} = 0$, with a misclassification error rate given by

$$\mathrm{ER}(\alpha) = \pi_1 \int_{g(\mathbf{x},\alpha)\leq 0} p(\mathbf{x}|\theta_1) d\mathbf{x} + \pi_0 \int_{g(\mathbf{x},\alpha)>0} p(\mathbf{x}|\theta_0) d\mathbf{x} \; .$$

The boundary actually used is $g(\mathbf{x}, \hat{\alpha}) = 0$, with a misclassification error rate given by

$$\text{ER}(\hat{\alpha}) = \pi_1 \int_{g(\mathbf{x},\hat{\alpha})\leq 0} p(\mathbf{x}|\theta_1)d\mathbf{x} + \pi_0 \int_{g(\mathbf{x},\hat{\alpha})>0} p(\mathbf{x}|\theta_0)d\mathbf{x} \geq \text{ER}(\alpha) \ .$$

Under some regularity conditions, O'Neill (1980) proved that, given that $\sqrt{n}(\hat{\alpha} - \alpha) \sim \mathcal{AN}(\mathbf{0}, \Sigma(\hat{\alpha}))$, the distribution of the random variable $n(\text{ER}(\hat{\alpha}) - \text{ER}(\alpha))$ converges to the distribution of the random variable $\xi^T \mathbf{B} \xi$, say:

$$n(\text{ER}(\hat{\alpha}) - \text{ER}(\alpha)) \rightarrow \xi^T \mathbf{B} \xi \ \text{ in distribution},$$

where $\xi \sim \mathcal{N}(\mathbf{0}, \Sigma(\hat{\alpha}))$, and

$$\mathbf{B} = \frac{1}{4} \int_D |\nabla_{\mathbf{x}} g(\mathbf{x}, \alpha)|^{-1} [\nabla_\alpha g(\mathbf{x}, \alpha)][\nabla_\alpha g(\mathbf{x}, \alpha)]^T p(\mathbf{x}) dm_D \ , \tag{3.1}$$

in which $D = \{\mathbf{x} : g(\mathbf{x}, \alpha) = 0\}$, m_D is Lebesgue measure on D, ∇_α and $\nabla_{\mathbf{x}}$ are vector partial differential operators corresponding to differentiation with respect to α and $\mathbf{x}$, $|\nabla_{\mathbf{x}} g(\mathbf{x}, \alpha)|$ is the L2-norm (also termed the Euclidean norm) of the vector $\nabla_{\mathbf{x}} g(\mathbf{x}, \alpha)$, and $p(\mathbf{x}) = \pi_1 p(\mathbf{x}|\theta_1) + \pi_0 p(\mathbf{x}|\theta_0)$.

Subsequently, Efron (1975) and O'Neill (1980) defined the asymptotic error rate (AER) as

$$\text{AER}(\hat{\alpha}) = \lim_{n\to\infty} E\{n(\text{ER}(\hat{\alpha}) - \text{ER}(\alpha))\} \ ,$$

which can be rewritten as

$$\text{AER}(\hat{\alpha}) = E\{\xi^T \mathbf{B} \xi\} = \text{tr}(E\{\xi^T \mathbf{B} \xi\}) = \text{tr}(\mathbf{B} E\{\xi \xi^T\}) = \text{tr}(\mathbf{B}\Sigma(\hat{\alpha})).$$

Since $\text{ER}(\hat{\alpha}) \geq \text{ER}(\alpha)$, the AER is actually a measure of an increased error rate because the estimated boundary is different from the optimal boundary.

Furthermore, Efron (1975) and O'Neill (1980) defined the ARE between two learning techniques as, for example,

$$\text{ARE}(\hat{\alpha}_{\text{d}}, \hat{\alpha}_{\text{g}}) = \frac{\text{AER}(\hat{\alpha}_{\text{g}})}{\text{AER}(\hat{\alpha}_{\text{d}})} = \frac{\text{tr}(\mathbf{B}\Sigma_{\text{g}}(\hat{\alpha}))}{\text{tr}(\mathbf{B}\Sigma_{\text{d}}(\hat{\alpha}))} \ .$$

If $\text{ARE}(\hat{\alpha}_{\text{d}}, \hat{\alpha}_{\text{g}}) < 1$, then generative learning provides estimators $\hat{\alpha}_{\text{g}}$ with lower asymptotic error rate with regard to the optimal discrimination coefficient α, *i.e.*, with less asymptotic misclassification error, than does discriminative learning; if $\text{ARE}(\hat{\alpha}_{\text{d}}, \hat{\alpha}_{\text{g}}) > 1$, then the relative performance of these two techniques reverses.

3.2.2 Theoretical Calculation of ARE

To calculate ARE for the discriminative, generative and GDT approaches, we need first to obtain $\mathbf{B}$, $\Sigma_{\mathrm{d}}(\hat{\alpha})$, $\Sigma_{\mathrm{g}}(\hat{\alpha})$ and $\Sigma_{\lambda}(\hat{\alpha})$.

For discriminative learning of the LLR estimator $\hat{\alpha}$, its asymptotic variance matrix $\Sigma_{\mathrm{d}}(\hat{\alpha})$ was proved by O'Neill (1980) to be

$$\Sigma_{\mathrm{d}}^{-1}(\hat{\alpha}) = \int_{\mathbf{x}} p(\mathcal{C}_1|\mathbf{x})p(\mathcal{C}_0|\mathbf{x})[\nabla_{\alpha} g(\mathbf{x},\alpha)][\nabla_{\alpha} g(\mathbf{x},\alpha)]^T p(\mathbf{x}) d\mathbf{x} \,. \tag{3.2}$$

It follows that, given $g(\mathbf{x},\alpha)$ (as in Equations (1.2) and (1.3) for linear and quadratic normal discrimination) and based on Equations (1.1), (3.1) and (3.2), $\mathbf{B}$ and $\Sigma_{\mathrm{d}}(\hat{\alpha})$ can be obtained.

As mentioned in Section 1.1.4, in order to obtain $\Sigma_{\mathrm{g}}(\hat{\alpha})$ and $\Sigma_{\lambda}(\hat{\alpha})$, we need first to derive $\Sigma_{\mathrm{g}}(\hat{\theta})$, $\Sigma_{\lambda}(\hat{\theta})$ and the relationship between $d\alpha = (\hat{\alpha} - \alpha)$ and $d\theta = (\hat{\theta} - \theta)$.

Asymptotic properties of maximum likelihood estimators suggest the following results.

First, $\sqrt{n}(\hat{\theta} - \theta) \sim \mathcal{AN}(\mathbf{0}, \Sigma_{\mathrm{g}}(\hat{\theta}) = nI_{\mathrm{g}}^{-1}(\theta))$, where $I_{\mathrm{g}}(\theta)$ is the Fisher information matrix,

$$I_{\mathrm{g}}(\theta) = E\left\{\frac{\partial \ell_{\mathrm{g}}(\theta)}{\partial \theta}\frac{\partial \ell_{\mathrm{g}}(\theta)}{\partial \theta^T}\right\} = E\left\{-\frac{\partial^2 \ell_{\mathrm{g}}(\theta)}{\partial \theta \partial \theta^T}\right\} \,.$$

Secondly,

$$\sqrt{n}(\hat{\theta} - \theta) \simeq \sqrt{n}\left[E\left\{-\frac{\partial^2 \ell_{\lambda}(\theta)}{\partial \theta \partial \theta^T}\right\}\right]^{-1} \cdot \frac{\partial \ell_{\lambda}(\theta)}{\partial \theta} \sim \mathcal{AN}(\mathbf{0}, \Sigma_{\lambda}(\hat{\theta})) \,,$$

where $\ell_{\lambda}(\theta) = \lambda \ell_{\mathrm{g}}(\theta) + (1-\lambda)\ell_{y|\mathbf{x}}(\theta)$, and $\Sigma_{\lambda}(\hat{\theta}) = nI_{\lambda}^{-1}(\theta)V_{\lambda}(\theta)I_{\lambda}^{-1}(\theta)$, in which, since $E\left\{\frac{\partial \ell_{\lambda}(\theta)}{\partial \theta}\right\} = 0$ and $\ell_{\mathrm{g}}(\theta) = \ell_{y|\mathbf{x}}(\theta) + \ell_{\mathbf{x}}(\theta)$,

$$I_{\lambda}(\theta) = E\left\{-\frac{\partial^2 \ell_{\lambda}(\theta)}{\partial \theta \partial \theta^T}\right\} = \lambda I_{\mathrm{g}}(\theta) + (1-\lambda) I_{y|\mathbf{x}}(\theta) \,,$$

$$V_{\lambda}(\theta) = \mathrm{Cov}\left(\frac{\partial \ell_{\lambda}(\theta)}{\partial \theta}\right) = E\left\{\left(\frac{\partial \ell_{\lambda}(\theta)}{\partial \theta}\right)^2\right\} = \lambda^2 I_{\mathrm{g}}(\theta) + (1-\lambda^2) I_{y|\mathbf{x}}(\theta) \,.$$

After some algebra, we obtain

$$\frac{1}{n} I_{y|\mathbf{x}}(\theta) = \int_{\mathbf{x}} p(\mathcal{C}_1|\mathbf{x})p(\mathcal{C}_0|\mathbf{x})\left[\frac{\partial \log r(\theta,\pi;\mathbf{x})}{\partial \theta}\right]\left[\frac{\partial \log r(\theta,\pi;\mathbf{x})}{\partial \theta}\right]^T p(\mathbf{x}) d\mathbf{x} \,,$$

with $r(\theta,\pi;\mathbf{x}) = \frac{\pi_1 p(\mathbf{x}|\theta_1)}{\pi_0 p(\mathbf{x}|\theta_0)}$ and $p(\mathbf{x}) = \pi_1 p(\mathbf{x}|\theta_1) + \pi_0 p(\mathbf{x}|\theta_0)$.

Meanwhile, based on a $g(\mathbf{x},\alpha)$ such as those defined in Equations (1.2) and (1.3) for linear and quadratic normal discrimination, we can obtain $d\alpha = M d\theta$ and thus $\Sigma_{\mathrm{g}}(\hat{\alpha}) = M\Sigma_{\mathrm{g}}(\hat{\theta})M^T$ and $\Sigma_{\lambda}(\hat{\alpha}) = M\Sigma_{\lambda}(\hat{\theta})M^T$.

Since a linear transformation of $\mathbf{x}$ into $a + A\mathbf{x}$ does not change the misclassification error rates, the above-mentioned calculation of asymptotic variance matrices can be simplified by a workable transformation. For example, for linear normal discrimination with $\mathbf{x}|\theta_1 \sim \mathcal{N}(\mu_1, \Sigma)$ and $\mathbf{x}|\theta_0 \sim \mathcal{N}(\mu_0, \Sigma)$, Efron (1975) suggested a new, linearly transformed $\mathbf{x}$ satisfying $\mathbf{x}|\theta_1 \sim \mathcal{N}(\frac{\Delta}{2}\mathbf{e}_1, \mathbf{I})$, $\mathbf{x}|\theta_0 \sim \mathcal{N}(-\frac{\Delta}{2}\mathbf{e}_1, \mathbf{I})$, where $\Delta = \sqrt{(\mu_1 - \mu_0)^T \Sigma^{-1} (\mu_1 - \mu_0)}$, the Mahalanobis distance between the means of the two sub-populations, and, in addition, it is required that $\Delta \neq 0$ to make the two sub-populations nonidentical; $\mathbf{I}$ is the identity matrix and $\mathbf{e}_1^T = (1, 0, 0, \ldots, 0)$. Another example is a linear transformation suggested by O'Neill (1980) for univariate quadratic normal discrimination.

The details of theoretical calculations and numerical evaluation of ARE for linear and quadratic normal discrimination can be found in the appendices of this thesis, as well as the corresponding details for the two examples suggested by Efron (1975) and O'Neill (1980), respectively.

3.2.3 Numerical Evaluations of ARE for Linear Normal Discrimination

The ARE between two learning techniques, with regard to estimators $\hat{\alpha}_1$ and $\hat{\alpha}_2$ of the coefficients of the discriminant function, is defined in Section 3.2.1 as $\text{ARE}(\hat{\alpha}_2, \hat{\alpha}_1) = \text{tr}(\mathbf{B}\Sigma(\hat{\alpha}_1))/\text{tr}(\mathbf{B}\Sigma(\hat{\alpha}_2))$.

For the example suggested by Efron (1975), theoretical derivation suggests that $\Sigma_{\text{g}}(\hat{\alpha})$, $\Sigma_{\lambda}(\hat{\alpha})$, $\Sigma_{\text{d}}(\hat{\alpha})$ and $\mathbf{B}$ are all symmetric block-diagonal matrices, represented by

$$\Sigma(\hat{\alpha}) = \begin{bmatrix} \Sigma_{1,1}^{(\hat{\alpha})} & \Sigma_{1,2}^{(\hat{\alpha})} & \\ \Sigma_{1,2}^{(\hat{\alpha})} & \Sigma_{2,2}^{(\hat{\alpha})} & \\ & & \Sigma_{3,3}^{(\hat{\alpha})}\mathbf{I}_{p-1} \end{bmatrix}, \; \mathbf{B} = \frac{\pi_1 \phi(\tau - \frac{\Delta}{2})}{2\Delta} \begin{bmatrix} 1 & \tau & \\ \tau & \tau^2 & \\ & & \mathbf{I}_{p-1} \end{bmatrix},$$

where $\phi(\cdot)$ denotes the density of the univariate standard normal distribution, p is the dimension of $\mathbf{x}$ and $\tau = -\frac{1}{\Delta}\log\frac{\pi_1}{\pi_0}$. It follows that

$$\begin{aligned} \text{tr}(\mathbf{B}\Sigma(\hat{\alpha})) &= \frac{\pi_1 \phi(\tau - \frac{\Delta}{2})}{2\Delta} \left\{ \text{tr}\left(\begin{bmatrix} 1 & \tau \\ \tau & \tau^2 \end{bmatrix} \begin{bmatrix} \Sigma_{1,1}^{(\hat{\alpha})} & \Sigma_{1,2}^{(\hat{\alpha})} \\ \Sigma_{1,2}^{(\hat{\alpha})} & \Sigma_{2,2}^{(\hat{\alpha})} \end{bmatrix} \right) + \text{tr}(\Sigma_{3,3}^{(\hat{\alpha})}\mathbf{I}_{p-1}) \right\} \\ &= \frac{\pi_1 \phi(\tau - \frac{\Delta}{2})}{2\Delta} \left\{ \Sigma_{1,1}^{(\hat{\alpha})} + 2\Sigma_{1,2}^{(\hat{\alpha})}\tau + \Sigma_{2,2}^{(\hat{\alpha})}\tau^2 + (p-1)\Sigma_{3,3}^{(\hat{\alpha})} \right\}. \end{aligned}$$

Therefore,

$$\text{ARE}(\hat{\alpha}_2, \hat{\alpha}_1) = \frac{\text{tr}(\mathbf{B}\Sigma(\hat{\alpha}_1))}{\text{tr}(\mathbf{B}\Sigma(\hat{\alpha}_2))} = \frac{\Sigma_{1,1}^{(\hat{\alpha}_1)} + 2\Sigma_{1,2}^{(\hat{\alpha}_1)}\tau + \Sigma_{2,2}^{(\hat{\alpha}_1)}\tau^2 + (p-1)\Sigma_{3,3}^{(\hat{\alpha}_1)}}{\Sigma_{1,1}^{(\hat{\alpha}_2)} + 2\Sigma_{1,2}^{(\hat{\alpha}_2)}\tau + \Sigma_{2,2}^{(\hat{\alpha}_2)}\tau^2 + (p-1)\Sigma_{3,3}^{(\hat{\alpha}_2)}}.$$

Here we present numerical evaluations of ARE as an index of comparison between the generative, discriminative and GDT approaches, for the case of linear normal discrimination under conditions (1) $\mathbf{x}|\theta_1 \sim \mathcal{N}(\frac{\Delta}{2}\mathbf{e}_1, \mathbf{I})$, $\mathbf{x}|\theta_0 \sim \mathcal{N}(-\frac{\Delta}{2}\mathbf{e}_1, \mathbf{I})$, (2) $\Delta \in [0.25, 4.75]$, (3) $\pi_1 \in [0.05, 0.95]$ and (4) $\lambda \in [0, 1]$.

3.2.3.1 Discriminative vs. Generative

Efron (1975) represented $\text{ARE}(\hat{\alpha}_{\text{d}}, \hat{\alpha}_{\text{g}})$ in terms of

$$Q_1 = \pi_1\pi_0 \left\{ [\Sigma_{\text{g}}(\hat{\alpha})]_{1,1} + 2[\Sigma_{\text{g}}(\hat{\alpha})]_{1,2}\tau + [\Sigma_{\text{g}}(\hat{\alpha})]_{2,2}\tau^2 \right\},$$

$$Q_2 = \pi_1\pi_0 \left\{ [\Sigma_{\text{g}}(\hat{\alpha})]_{3,3} \right\},$$

$$Q_3 = \pi_1\pi_0 \left\{ [\Sigma_{\text{d}}(\hat{\alpha})]_{1,1} + 2[\Sigma_{\text{d}}(\hat{\alpha})]_{1,2}\tau + [\Sigma_{\text{d}}(\hat{\alpha})]_{2,2}\tau^2 \right\},$$

$$Q_4 = \pi_1\pi_0 \left\{ [\Sigma_{\text{d}}(\hat{\alpha})]_{3,3} \right\},$$

$$\text{Eff}_{p=1} = Q_1/Q_3,\ \text{Eff}_{p\to\infty} = Q_2/Q_4,$$

and hence

$$\text{Eff}_p = \text{ARE}(\hat{\alpha}_{\text{d}}, \hat{\alpha}_{\text{g}}) = \frac{Q_1 + (p-1)Q_2}{Q_3 + (p-1)Q_4} = \frac{\frac{Q_3}{Q_4}\text{Eff}_{p=1} + (p-1)\text{Eff}_{p\to\infty}}{\frac{Q_3}{Q_4} + (p-1)}.$$

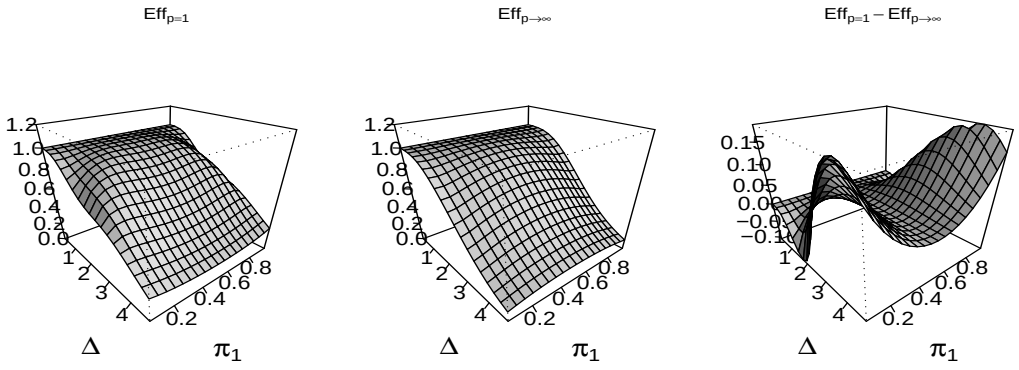


Figure 3.1: The ARE between the generative approach and the discriminative approach for linear normal discrimination: left-hand panel gives $\text{Eff}_{p=1}$, middle panel gives $\text{Eff}_{p\to\infty}$, right-hand panel gives $\text{Eff}_{p=1} - \text{Eff}_{p\to\infty}$.

Numerical evaluations of $\text{Eff}_{p=1}$, $\text{Eff}_{p\to\infty}$ and their difference are shown in Figure 3.1. We make the following observations.

1. Both $\text{Eff}_{p=1}$ and $\text{Eff}_{p\to\infty}$ are less than 1, indicating that asymptotically the generative approach will provide better classification accuracy than the discriminative approach.

2. Both $\text{Eff}_{p=1}$ and $\text{Eff}_{p\to\infty}$ decrease as the Mahalanobis distance Δ increases; this implies that, for two well-separated sub-populations, the generative approach is much better than the discriminative approach; in other words, the latter may be an acceptable alternative to the former only when the two sub-populations are poorly separated, with $\Delta < 2$.

3. Sometimes $\text{Eff}_{p=1}$ can be smaller than $\text{Eff}_{p\to\infty}$; however, in agreement with Efron (1975), it is more likely that $\text{Eff}_{p=1} \geq \text{Eff}_{p\to\infty}$; this implies that, when we use the discriminative approach as an alternative to the generative approach for high-dimensional data, it is more likely to lower the classification accuracy, in particular when the Mahalanobis distance $\Delta > 2$.

3.2.3.2 GDT vs. Generative

Similarly, we define Q_5 and Q_6 by

$$Q_5 = \pi_1\pi_0 \left\{ [\Sigma_\lambda(\hat{\alpha})]_{1,1} + 2[\Sigma_\lambda(\hat{\alpha})]_{1,2}\tau + [\Sigma_\lambda(\hat{\alpha})]_{2,2}\tau^2 \right\},$$

$$Q_6 = \pi_1\pi_0 \left\{ [\Sigma_\lambda(\hat{\alpha})]_{3,3} \right\} ,$$

so that

$$\text{Eff}^{(\lambda)}_{p=1} = Q_1/Q_5,\ \text{Eff}^{(\lambda)}_{p\to\infty} = Q_2/Q_6 ,$$

and hence

$$\text{Eff}^{(\lambda)}_{p} = \text{ARE}(\hat{\alpha}_\lambda, \hat{\alpha}_\text{g}) = \frac{Q_1 + (p-1)Q_2}{Q_5 + (p-1)Q_6} = \frac{\frac{Q_5}{Q_6}\text{Eff}^{(\lambda)}_{p=1} + (p-1)\text{Eff}^{(\lambda)}_{p\to\infty}}{\frac{Q_5}{Q_6} + (p-1)} .$$

Numerical evaluations of $\text{Eff}^{(\lambda)}_{p=1}$, $\text{Eff}^{(\lambda)}_{p\to\infty}$ and their difference are shown in Figure 3.2, for $\lambda = 0, 0.25, 0.5$ and 0.75, respectively. We make the following observations.

1. For all these values of λ, both $\text{Eff}^{(\lambda)}_{p=1}$ and $\text{Eff}^{(\lambda)}_{p\to\infty}$ are less than 1, indicating that asymptotically the generative approach will provide better classification accuracy than the GDT.

2. When $\lambda = 0$, the GDT contains its discriminative component alone. For such a case, similarly to the ARE between the discriminative approach and the generative approach

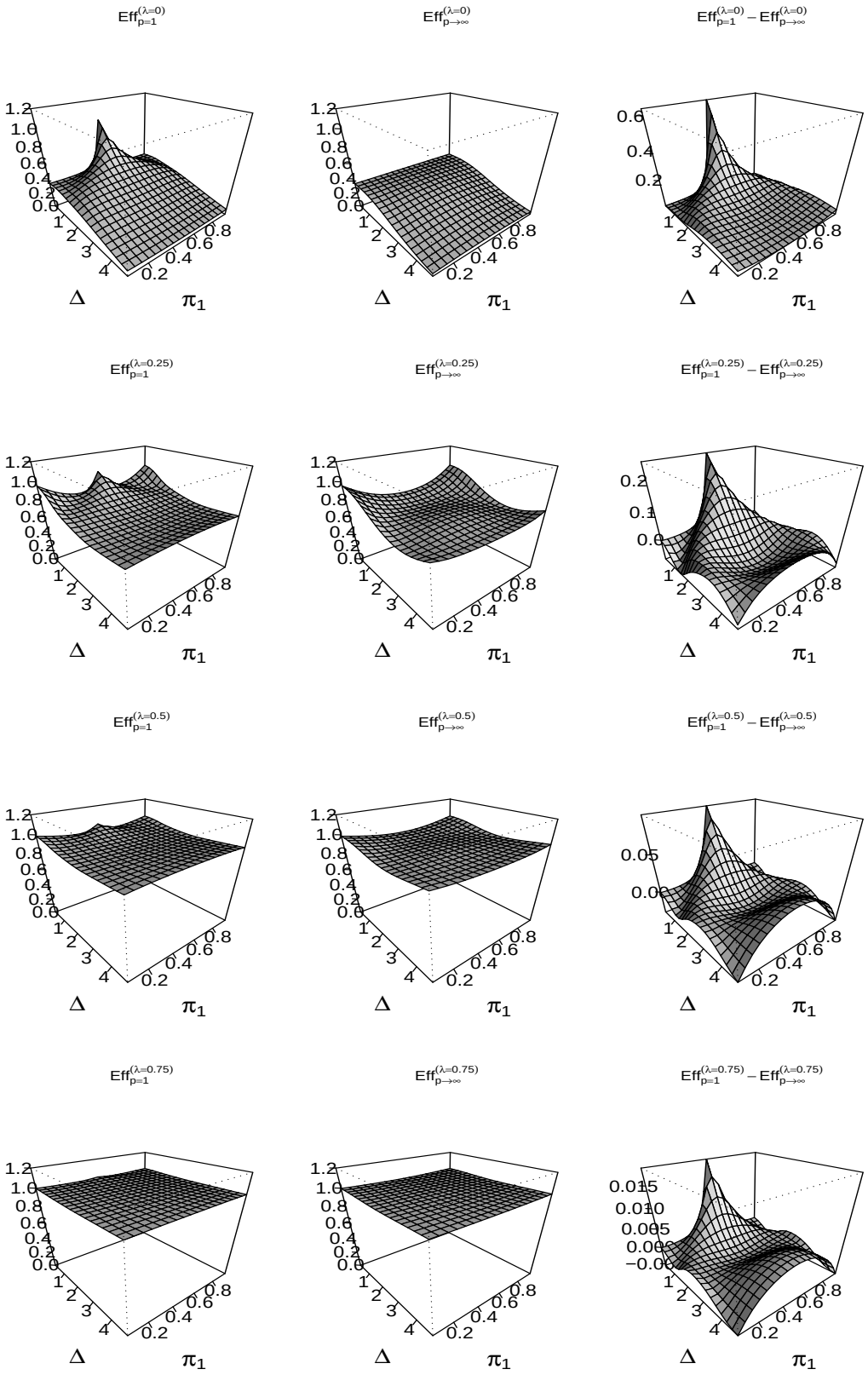


Figure 3.2: The ARE between the generative approach and the GDT with $\lambda = 0, 0.25, 0.5$ and 0.75, respectively, for linear normal discrimination: first column gives $\text{Eff}_{p=1}^{(\lambda)}$, second column gives $\text{Eff}_{p\to\infty}^{(\lambda)}$, third column gives $\text{Eff}_{p=1}^{(\lambda)} - \text{Eff}_{p\to\infty}^{(\lambda)}$.

(as shown in Figure 3.1), both $\text{Eff}_{p=1}^{(\lambda=0)}$ and $\text{Eff}_{p\to\infty}^{(\lambda=0)}$ in general decrease as the Mahalanobis distance Δ increases. This implies bad classification accuracy of the GDT with $\lambda = 0$.

However, this is not the case for other values of λ, where both $\text{Eff}_{p=1}^{(\lambda)}$ and $\text{Eff}_{p\to\infty}^{(\lambda)}$ fluctuate as Δ increases, since the GDT contains a generative component. The minima of $\text{Eff}_{p=1}^{(\lambda=0.25)}$ and $\text{Eff}_{p\to\infty}^{(\lambda=0.25)}$ are both larger than 0.64. This implies that, even though the generative component only has a small weight, the GDT can act as an acceptable alternative to the generative approach.

3. When $\lambda = 0$, we have $\text{Eff}_{p=1}^{(\lambda)} \geq \text{Eff}_{p\to\infty}^{(\lambda)}$; for other values of λ, this inequality usually holds for most settings of Δ and π_1. This implies that, when the GDT is used as an alternative to the generative approach for high-dimensional data, it usually lowers the classification accuracy.

4. Apparently, when λ increases so that the generative component of the GDT gains more weight, then the ARE, namely $\text{Eff}_p^{(\lambda)}$, is closer to 1, in which case the GDT equates to the generative approach.

3.2.3.3 Discriminative vs. GDT

The ARE between the discriminative approach and the GDT is simply the ratio of the AREs between them and the generative approach, described in Sections 3.2.3.1 and 3.2.3.2. That is,

$$\text{ARE}(\hat{\alpha}_{\text{d}}, \hat{\alpha}_{\lambda}) = \frac{\text{ARE}(\hat{\alpha}_{\text{d}}, \hat{\alpha}_{\text{g}})}{\text{ARE}(\hat{\alpha}_{\lambda}, \hat{\alpha}_{\text{g}})} = \frac{\text{Eff}_p}{\text{Eff}_p^{(\lambda)}} = \frac{Q_5 + (p-1)Q_6}{Q_3 + (p-1)Q_4} = \frac{\frac{Q_3}{Q_4}\frac{\text{Eff}_{p=1}}{\text{Eff}_{p=1}^{(\lambda)}} + (p-1)\frac{\text{Eff}_{p\to\infty}}{\text{Eff}_{p\to\infty}^{(\lambda)}}}{\frac{Q_3}{Q_4} + (p-1)},$$

where

$$\frac{\text{Eff}_{p=1}}{\text{Eff}_{p=1}^{(\lambda)}} = \frac{Q_5}{Q_3}, \ \frac{\text{Eff}_{p\to\infty}}{\text{Eff}_{p\to\infty}^{(\lambda)}} = \frac{Q_6}{Q_4}.$$

If $\text{ARE}(\hat{\alpha}_{\text{d}}, \hat{\alpha}_{\lambda}) < 1$, then the GDT performs better than the discriminative approach, in terms of the asymptotic misclassification error; if $\text{ARE}(\hat{\alpha}_{\text{d}}, \hat{\alpha}_{\lambda}) > 1$, then the discriminative approach performs better.

Lemma 3.2.1 *When $\lambda = 1$, we have $ARE(\hat{\alpha}_d, \hat{\alpha}_\lambda) = Eff_p$; When $\lambda = 0$, we have $\frac{Eff_{p\to\infty}}{Eff_{p\to\infty}^{(\lambda=0)}} = \frac{[\Sigma_\lambda(\hat{\alpha})]_{3,3}}{[\Sigma_d(\hat{\alpha})]_{3,3}} \equiv 3$.* ∎

Lemma 3.2.1 shows that the ARE between the discriminative approach and the GDT with $\lambda = 0$ converges to 3 when $p \to \infty$. This implies that, for high-dimensional data, the discriminative approach, compared to the GDT's discriminative component, converges to a threefold improvement in the classification performance as measured by the misclassification error rate. In addition, after some algebra, we have the following lemma, which implies that, for balanced data, a discriminative approach is favoured, rather than the GDT's discriminative component.

Lemma 3.2.2 *When $\pi_1 = \pi_0 = \frac{1}{2}$, we have*

$$\frac{\mathit{Eff}_p}{\mathit{Eff}_p^{(\lambda=0)}} = \frac{[\Sigma_\lambda(\hat{\alpha})]_{1,1} + (p-1)[\Sigma_\lambda(\hat{\alpha})]_{3,3}}{p[\Sigma_d(\hat{\alpha})]_{3,3}}$$

$$= \frac{1}{p}\left(1 + \frac{2\Delta^2 A_0}{4A_2 + \Delta^2 A_0}\right) + \frac{3(p-1)}{p} = \frac{1}{p}\left(\frac{-8A_2}{4A_2 + \Delta^2 A_0}\right) + 3 \geq 1. \quad \blacksquare$$

Numerical evaluations of $\frac{\text{Eff}_{p=1}}{\text{Eff}_{p=1}^{(\lambda)}}$, $\frac{\text{Eff}_{p\to\infty}}{\text{Eff}_{p\to\infty}^{(\lambda)}}$ and their difference are shown in Figure 3.3, for $\lambda = 0, 0.25, 0.5$ and 0.75, respectively. We make the following observations.

1. When $\lambda = 0$, $\frac{\text{Eff}_{p=1}}{\text{Eff}_{p=1}^{(\lambda=0)}} \geq 1$ and $\frac{\text{Eff}_{p\to\infty}}{\text{Eff}_{p\to\infty}^{(\lambda=0)}} \equiv 3$, indicating that asymptotically the discriminative approach will provide better classification accuracy than the GDT's discriminative component alone. However, as λ increases, both $\frac{\text{Eff}_{p=1}}{\text{Eff}_{p=1}^{(\lambda)}}$ and $\frac{\text{Eff}_{p\to\infty}}{\text{Eff}_{p\to\infty}^{(\lambda)}}$ reduce in value to be less than 1 for increasingly many settings of Δ and π_1, indicating a reverse of the relative performance of the two approaches.

2. When $\lambda = 0$, it is more likely that $\frac{\text{Eff}_{p=1}}{\text{Eff}_{p=1}^{(\lambda=0)}} < \frac{\text{Eff}_{p\to\infty}}{\text{Eff}_{p\to\infty}^{(\lambda=0)}}$, while, for other values of λ, it is more likely that $\frac{\text{Eff}_{p=1}}{\text{Eff}_{p=1}^{(\lambda)}} > \frac{\text{Eff}_{p\to\infty}}{\text{Eff}_{p\to\infty}^{(\lambda)}}$.

3. Apparently, when λ increases so that the GDT's generative component gains more weight, then the ARE, namely $\frac{\text{Eff}_p}{\text{Eff}_p^{(\lambda)}}$, approaches Eff_p (as shown in Figure 3.1), as $\text{Eff}_p^{(\lambda)} \to 1$.

3.3 Simulation Study on Classification Performance of GDT

3.3.1 Implementation

The hybrid learning can be viewed as an optimisation problem for multi-classifiers. The optimisation of the GDT is based on a new log-likelihood, $\ell_\lambda(\theta)$, based on the common parameter vector θ. Here, for generalisation to the case of multi-groups, we re-write $\ell_\lambda(\theta)$ as

$$\ell_\lambda(\theta) = \lambda\ell_{\text{g}}(\theta) + (1-\lambda)\ell_{y|\mathbf{x}}(\theta) = \ell_{\text{g}}(\theta) - (1-\lambda)\ell_{\mathbf{x}}(\theta)\ , \text{ with}$$

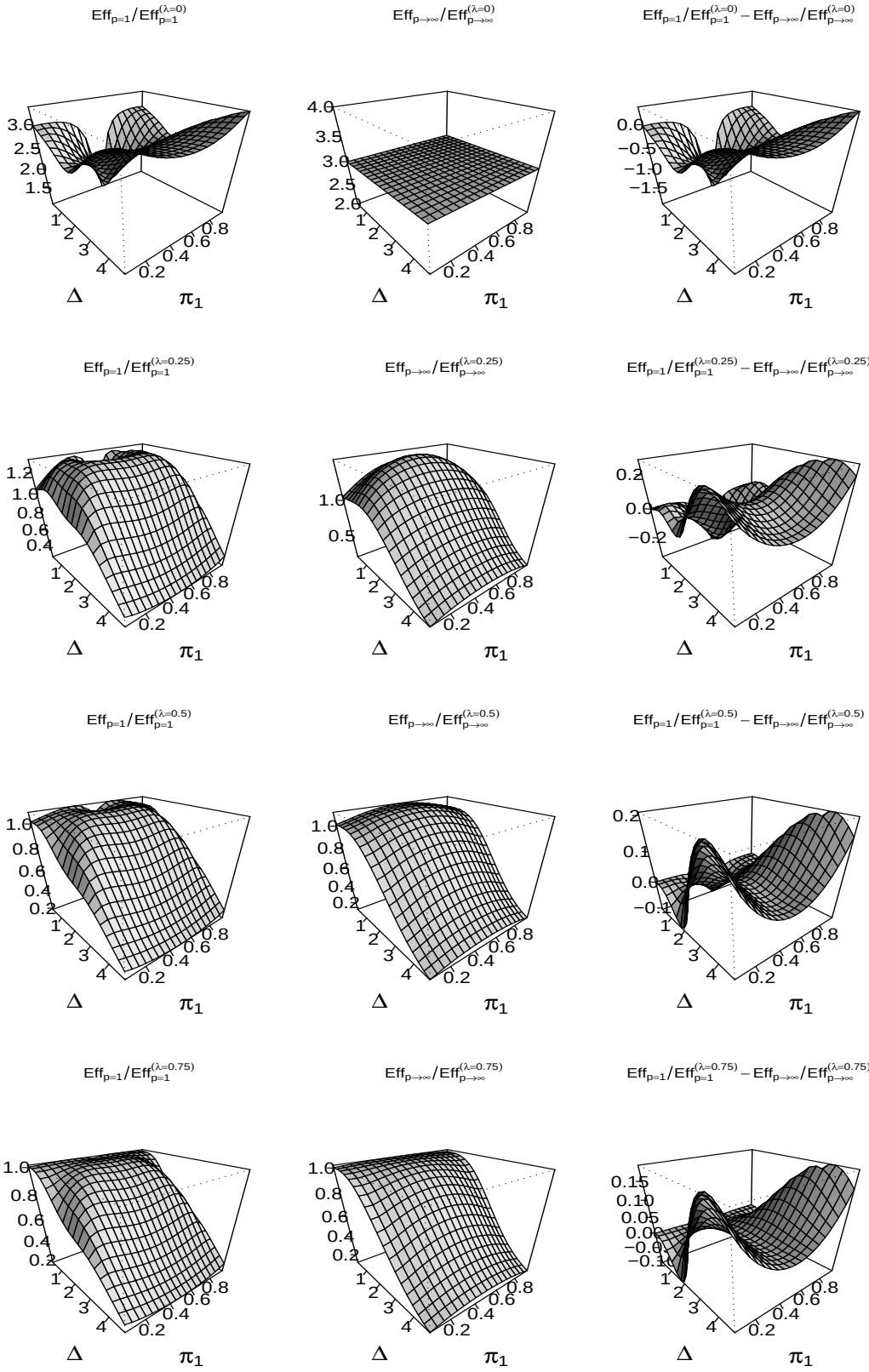


Figure 3.3: The ARE between the GDT and the discriminative approach with $\lambda = 0, 0.25, 0.5$ and 0.75, respectively, for linear normal discrimination: first column gives $\frac{\mathrm{Eff}_{p=1}}{\mathrm{Eff}_{p=1}^{(\lambda)}}$, second column gives $\frac{\mathrm{Eff}_{p\to\infty}}{\mathrm{Eff}_{p\to\infty}^{(\lambda)}}$, third column gives $\frac{\mathrm{Eff}_{p=1}}{\mathrm{Eff}_{p=1}^{(\lambda)}} - \frac{\mathrm{Eff}_{p\to\infty}}{\mathrm{Eff}_{p\to\infty}^{(\lambda)}}$.

$$\ell_{\mathrm{g}}(\theta) = \log \prod_{i=1}^{n} p(\mathbf{x}_i, y_i; \theta) = \log \prod_{i=1}^{n} \pi_{y_i} p(\mathbf{x}_i | y_i; \theta_{y_i}) \ ,$$

$$\ell_{y|\mathbf{x}}(\theta) = \log \prod_{i=1}^{n} p(y_i | \mathbf{x}_i) = \log \prod_{i=1}^{n} \frac{\pi_{y_i} p(\mathbf{x}_i | y_i; \theta_{y_i})}{\sum_{k=1}^{K} \pi_k p(\mathbf{x}_i | y = k; \theta_k)} \ ,$$

$$\ell_{\mathbf{x}}(\theta) = \log \prod_{i=1}^{n} p(\mathbf{x}_i) = \log \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} \pi_k p(\mathbf{x}_i | y = k; \theta_k) \right\} \ .$$

in which $\pi_{y_i} = p(y = y_i)$, $y_i \in \{1, \ldots, K\}$, K is the number of groups ($K = 2$ in our study), and θ consists of π_k and θ_k, $k = 1, \ldots, K$, a parameter vector of the joint distribution $p(\mathbf{x}, y; \theta)$. As seen from $\ell_\lambda(\theta)$, the GDT becomes a pure generative approach when $\lambda = 1$ while the weight of its discriminative component increases as λ decreases from 1 to 0.

We use a general-purpose optimization based on the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, a quasi-Newton method, implemented by an R function *optim* from the standard package **stats** in R. Meanwhile, in order to investigate the performance discrepancy between the discriminative component of the GDT and a truly discriminative approach, we compared the GDT at $\lambda = 0$ with LLR. Here LLR is implemented by an R function *logitreg* (Venables and Ripley, 2002), also using the BFGS algorithm.

In order to implement a GDT, the conditional distribution $p(\mathbf{x}|y)$ has to be specified; as was done in the simulation study by Bouchard and Triggs (2004), we assume that $(\mathbf{x}|y)$ follows multivariate normal distributions $\mathcal{N}(\mu_k, \Lambda)$ with a common diagonal covariance matrix Λ across the groups. However, we do not assume equal prior probabilities π_k but estimate them from the training samples instead.

For the assumed Gaussian model with a common diagonal covariance matrix Λ across the K groups, the parameter vector θ is composed of $K - 1$ prior probabilities $\{\pi_k\}_{k=1}^{K-1}$, K p-dimensional mean vectors $\{\mu_k\}_{k=1}^{K}$ and the p diagonal components $\{\Lambda_{j,j}\}_{j=1}^{p}$ of Λ.

First, the derivatives of $\ell_\lambda(\theta)$ with respect to $\{\pi_k\}_{k=1}^{K-1}$ can be written as

$$\frac{\partial \ell_\lambda(\theta)}{\partial \pi_k} = \sum_{i=1}^{n} \left\{ \frac{\mathbf{1}_{\{y_i = k\}} - (1 - \lambda) p(y = k | \mathbf{x}_i)}{\pi_k} - \frac{\mathbf{1}_{\{y_i = K\}} - (1 - \lambda) p(y = K | \mathbf{x}_i)}{\pi_K} \right\} \ ,$$

where, as in Bouchard and Triggs (2004),

$$p(y = k | \mathbf{x}_i) = \frac{\pi_k p(\mathbf{x}_i | y = k; \theta_k)}{\sum_{l=1}^{K} \pi_l p(\mathbf{x}_i | y = l; \theta_l)} \ .$$

Secondly, the derivatives of $\ell_\lambda(\theta)$ with respect to $\{\mu_k\}_{k=1}^{K}$, unique for each group, can be

written as

$$\frac{\partial \ell_\lambda(\theta)}{\partial \mu_k} = \sum_{i=1}^{n} \left\{ \left\{ \mathbf{1}_{\{y_i=k\}} - (1-\lambda) p(y=k|\mathbf{x}_i) \right\} \frac{\partial \log p(\mathbf{x}_i|y=k;\theta_k)}{\partial \mu_k} \right\} ,$$

where, for the assumed Gaussian model with Λ,

$$\frac{\partial \log p(\mathbf{x}_i|y=k;\theta_k)}{\partial \mu_k} = \Lambda^{-1}(\mathbf{x}_i - \mu_k) .$$

Thirdly, the derivatives of $\ell_\lambda(\theta)$ with respect to Λ, common for all the groups, can be written as

$$\frac{\partial \ell_\lambda(\theta)}{\partial \Lambda} = \sum_{k=1}^{K} \sum_{i=1}^{n} \left\{ \left\{ \mathbf{1}_{\{y_i=k\}} - (1-\lambda) p(y=k|\mathbf{x}_i) \right\} \frac{\partial \log p(\mathbf{x}_i|y=k;\theta_k)}{\partial \Lambda} \right\} ,$$

where, for the assumed Gaussian model with Λ,

$$\frac{\partial \log p(\mathbf{x}_i|y=k;\theta_k)}{\partial \Lambda} = \frac{1}{2} \left\{ -\Lambda^{-1} + \Lambda^{-1}(\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^T \Lambda^{-1} \right\} .$$

The above formulae can be rewritten with matrix representations so as to facilitate the computation by matrix-based software like Matlab and R. A simple example of this is that, if $p(y=k|\mathbf{x}_i)$ and, for each $\Lambda_{j,j}$, $\frac{\partial \log p(\mathbf{x}_i|y=k;\theta_k)}{\partial \Lambda_{j,j}}$ are assembled into two $K \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$, respectively, then

$$\sum_{k=1}^{K} \sum_{i=1}^{n} p(y=k|\mathbf{x}_i) \frac{\partial \log p(\mathbf{x}_i|y=k;\theta_k)}{\partial \Lambda_{j,j}} = \text{trace}(\mathbf{A}^T \mathbf{B}) .$$

In our study, four datasets are simulated; one of them, arising from two normal distributions with a common identity covariance matrix $\mathbf{I}$ which exactly satisfies the modelling assumptions about the data-generating process $p(\mathbf{x}|y)$, is also used by Bouchard and Triggs (2004), and the other three are all from two normal distributions but with either a common full covariance matrix or two unequal diagonal covariance matrices or two unequal full covariance matrices, respectively. All of the latter three datasets violate the modelling assumptions about $p(\mathbf{x}|y)$, and all the distributions are 4-dimensional, *i.e.*, all the data are of two groups with four features.

Meanwhile, in order to investigate how the classification performance depends on both the training-set size n and the weight λ, n is sampled within $[50, 250]$ in steps of 25, and λ is sampled within $[0, 1]$ in steps of 0.1; the test set size is 10^3 since at this size our results for the logistic loss are at a similar level to those reported in Bouchard and Triggs (2004). Within the range $[0, 1]$ of λ, we use the same optimisation procedure to estimate the parameter vector θ

with LDA-Λ equivalent to the GDT at $\lambda = 1$, while the results obtained from LLR are recorded and plotted at $\lambda = -0.1$ so as to be neighbours of those of the discriminative component of the GDT at $\lambda = 0$ for comparison only, where $\lambda = -0.1$ has no meaning in terms of physical weight. For each sampled n, the 10^3 observations are randomly split into n training samples and $10^3 - n$ test samples with 100 replicates; from them, the medians of the logistic losses and misclassification error rates are recorded and plotted. Sample proportions and moments are used as the initial values for BFGS optimisation.

Along with the logistic loss used by Bouchard and Triggs (2004), we also use the traditional misclassification error rate (ER) to measure the performance of the classifiers, defined as usual by the number of misclassified observations over the total number of observations for binary discrimination. For the dataset used by both Bouchard and Triggs (2004) and ourselves, our results about the logistic loss in general lead to similar observations to those reported in Bouchard and Triggs (2004). Therefore, in the following, we only report the results about the ER.

3.3.2 Normally Distributed Data

Four simulated datasets are used in this section, each consisting of 10^3 samples that are randomly generated from two 4-variate normal distributions, $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$, based on 500 samples from each distribution. As in Bouchard and Triggs (2004), $\mu_1 = (1.25, 0, 0, 0)^T$ and $\mu_2 = (-1.25, 0, 0, 0)^T$, where μ_2 only differs from μ_1 in one of the four dimensions; other values of μ_1 and μ_2 can be linearly transformed to these two values so that there is no loss of generality. Meanwhile, Σ_1 and Σ_2 are subject to four different types of constraint, specified as follows.

1. **Equal diagonal covariance matrices:** $\Sigma_1 = \Sigma_2 = \Lambda = \mathbf{I}$.
2. **Equal full covariance matrices:** $\Sigma_1 = \Sigma_2 = \Sigma$ while $\Sigma \neq \Lambda$, with $\Sigma = \begin{bmatrix} 1 & c & c & c \\ c & 1 & c & c \\ c & c & 1 & c \\ c & c & c & 1 \end{bmatrix}$ with $c = 0.25$.
3. **Unequal diagonal covariance matrices:** $\Sigma_1 = \Lambda_1$, $\Sigma_2 = \Lambda_2$ with $\Lambda_1 \neq \Lambda_2$, where $\Lambda_1 = \mathbf{I}$ and $\Lambda_2 = \text{Diag}(0.25, 0.75, 1.25, 1.75)$.

4. **Unequal full covariance matrices:** $\Sigma_1 \neq \Lambda_1$, $\Sigma_2 \neq \Lambda_2$ and $\Sigma_1 \neq \Sigma_2$, with $\Sigma_1 = \begin{bmatrix} 1 & c & c & c \\ c & 1 & c & c \\ c & c & 1 & c \\ c & c & c & 1 \end{bmatrix}$ and $\Sigma_2 = \begin{bmatrix} 0.25 & c & c & c \\ c & 0.75 & c & c \\ c & c & 1.25 & c \\ c & c & c & 1.75 \end{bmatrix}$ with $c = 0.25$.

3.3.3 Results

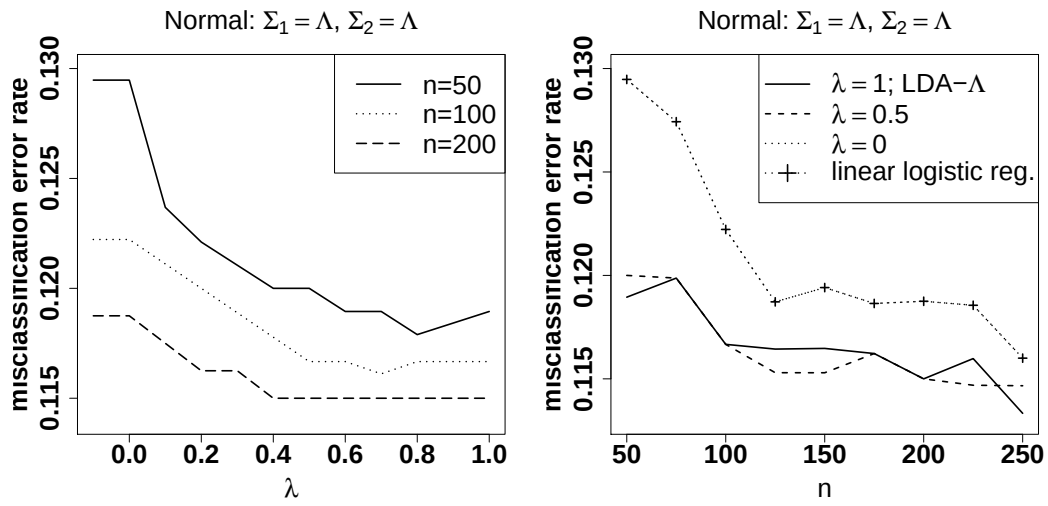


Figure 3.4: Simulated normally distributed data with equal diagonal covariance matrices. Plots of classification performance measured by ER vs. training-set size n and λ ($\lambda = -0.1$ corresponds to LLR, $\lambda \in [0, 1]$ corresponds to GDT and $\lambda = 1$ corresponds to LDA-Λ), obtained from 100 experiments on test sets of size 10^3. Left-hand panel: ER vs. λ for $n = 50, 100$ and 200; right-hand panel: ER vs. n for LDA-Λ, $\lambda = 0.5, 0$ and LLR.

Our results are shown in Figures 3.4, 3.5, 3.6 and 3.7, respectively, for the four simulated datasets. Each figure consists of two plots of the ER vs. λ and the ER vs. n, respectively; from them, we observe the following patterns.

1. For the first dataset in which no mis-specification of the assumed Gaussian model with Λ occurs except for there being a finite number of observations in the training set, as shown in Figure 3.4, LDA-Λ in general performs the best.

2. When there is mis-specification, such as those cases shown in Figures 3.5-3.7, at some optimal values of $\lambda \in (0, 1)$ the GDT can perform better than at $\lambda = 0$ and $\lambda = 1$.

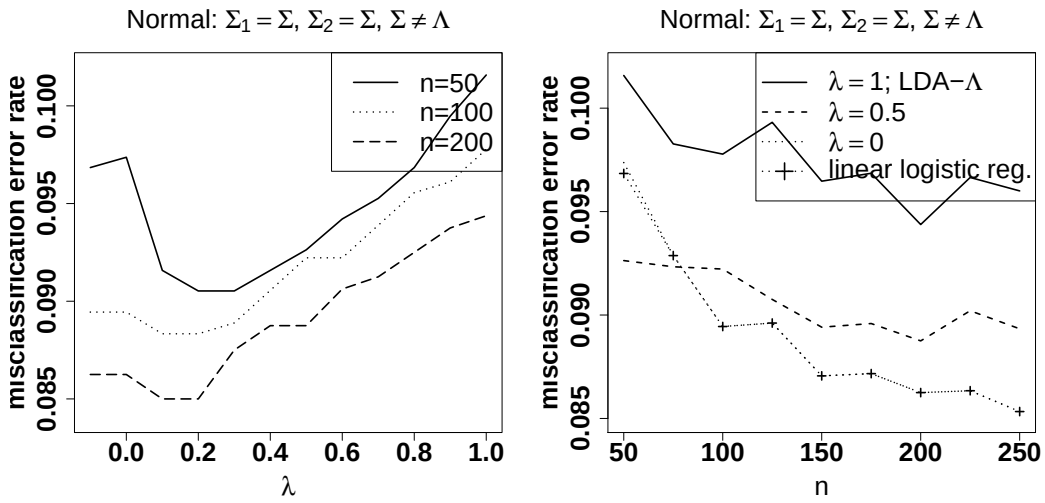


Figure 3.5: Simulated normally distributed data with equal full covariance matrices.

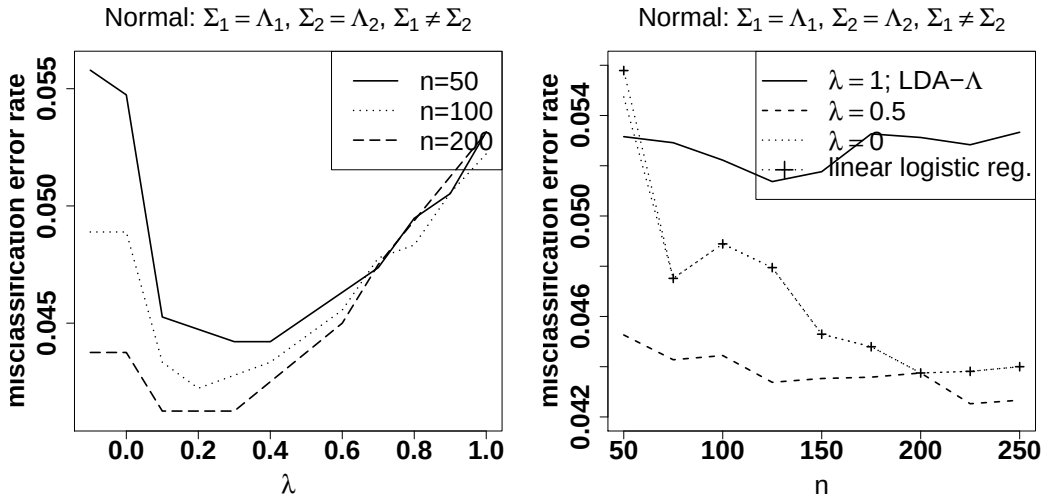


Figure 3.6: Simulated normally distributed data with unequal diagonal covariance matrices.

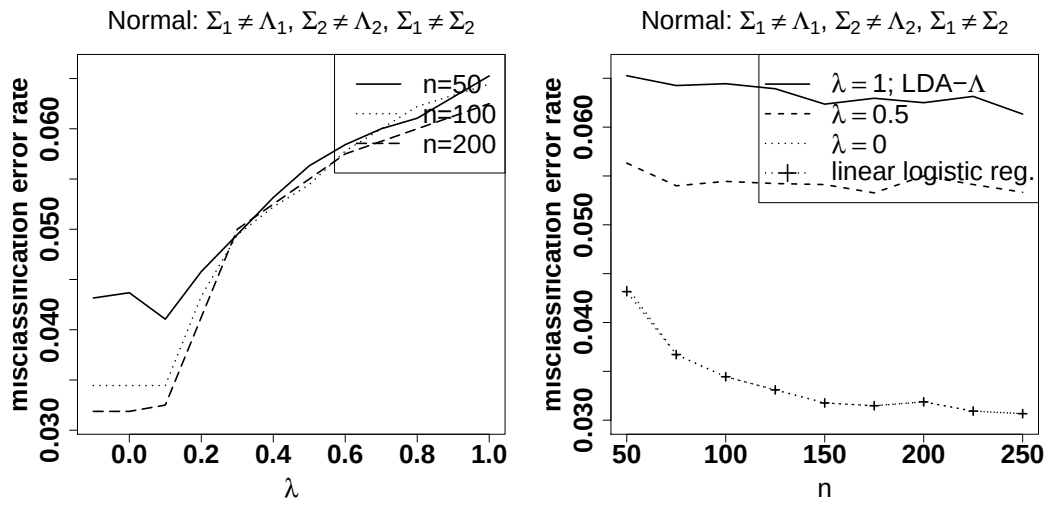


Figure 3.7: Simulated normally distributed data with unequal full covariance matrices.

3. When there is mis-specification and the training-set size n is large, our results show that the performance of LLR, a discriminative classifier, is superior to that of LDA-Λ, a generative one.

4. Our results support the claim made by Bouchard (2005) that, under our assumption of common diagonal covariance matrices, the discriminative component of the GDT (with $\lambda = 0$) performs the same as LLR does, as they optimise the same objective function. Nevertheless, our results also show that, when there is mis-specification and n is small, practical optimisation with regard to different parameterisations may either converge at different values or even stop iteration without convergence.

3.4 Conclusions

The conclusions from our study are three-fold.

First, the GDT is a generative model integrating both discriminative and generative learning, so that it is also subject to model mis-specification of the data-generating process $p(\mathbf{x}|y;\theta_{\mathrm{g}})$, or otherwise of the joint distribution $p(\mathbf{x}, y;\theta)$.

Secondly, amongst the three approaches that we compare, the asymptotic efficiency of the GDT is lower than that of generative learning when there is no model mis-specification.

Thirdly, when there is no model mis-specification, LDA performs the best; when there

is model mis-specification, the GDT may perform the best at an optimal tradeoff between its discriminative and generative components, and LLR, a truly discriminative classifier, in general performs well when the training-sample size n is reasonably large.

Chapter 4

On the Hybrid Generative/Discriminative Algorithm

The so-called hybrid generative/discriminative algorithm assigns different weights to partial feature vectors of $\mathbf{x}$, learning most parameters generatively but the weights discriminatively (Raina et al., 2003). In this chapter, we first interpret the hybrid algorithm from three perspectives, namely class-conditional probabilities, class-posterior probabilities and loss functions underlying the model, and then discuss one of its multi-class extensions (Fujino et al., 2007). Finally, by using simulated and real-world data, we compare its classification performance with that of the naïve Bayes classifier and linear logistic regression.

4.1 Interpretation of the Hybrid Algorithm

Consider classifying an observation with h features into one of K groups by a classifier $\hat{y}$, which was trained by using the observed features and group labels of m other so-called training observations. In this chapter, the dimension of features is denoted by h instead of p. We use an h-variate random vector $\mathbf{x} = (x_1, \ldots, x_h)^T$ to represent the h features of the observation and a random categorical variable $y \in \{1, \ldots, K\}$ to represent the group label. We denote a classifier of $\mathbf{x}$ by $\hat{y}(\mathbf{x})$ and the loss function of misclassifying $\mathbf{x}$, which arises from the group y, into the group $\hat{y}(\mathbf{x})$ is $L(y, \hat{y}(\mathbf{x}))$.

4.1.1 Class-conditional Probabilities

For binary classification, where $K = 2$, based on Bayes' Theorem, the Bayes discriminant criterion (*i.e.*, $\hat{y}(\mathbf{x}) = \operatorname{argmax}_y p(y|\mathbf{x})$) of the generative classifiers for classifying $\mathbf{x}$ into the group $y = 1$ can be written as $p(\mathbf{x}, y = 1) \geq p(\mathbf{x}, y = 2)$, or equivalently $p(y = 1)p(\mathbf{x}|y = 1) \geq p(y = 2)p(\mathbf{x}|y = 2)$. In addition, specific generative classifiers, such as linear normal-based discriminant analysis with a common diagonal covariance matrix (denoted by LDA-Λ) and the naïve Bayes classifier, assume that the h features are conditionally independent given the group label y, *i.e.*, $p(\mathbf{x}|y) = \prod_{i=1}^{h} p(x_i|y)$.

In the normalised hybrid and the unnormalised hybrid algorithms proposed by Raina et al. (2003), the feature vector $\mathbf{x}$ is divided into R partial feature vectors $\mathbf{x}^1, \ldots, \mathbf{x}^R$, because they suggest different levels of importance for different partitions, or partial feature vectors; for example, $\mathbf{x}^1$ may represent the message subject of an email while $\mathbf{x}^2$ represents the message body. As with Raina et al. (2003), we focus on $R = 2$, such that $\mathbf{x} = (\mathbf{x}^{1T}, \mathbf{x}^{2T})^T$, $\mathbf{x}^1 = (x_1, \ldots, x_{h_1})^T$, $\mathbf{x}^2 = (x_{h_1+1}, \ldots, x_h)^T$ and $h_2 = h - h_1$, and assume that the discriminant criterion of the generative classifiers can be rewritten as

$$p(y = 1)p(\mathbf{x}^1|y = 1)p(\mathbf{x}^2|y = 1) \geq p(y = 2)p(\mathbf{x}^1|y = 2)p(\mathbf{x}^2|y = 2) \, .$$

Thus, given $p(\mathbf{x}, y) \neq 0$, the corresponding discriminant function $\lambda_{\mathrm{G}}(\mathbf{x}) = \log \frac{p(y=1|\mathbf{x})}{p(y=2|\mathbf{x})}$ can be expressed in terms of likelihood ratios as

$$\lambda_{\mathrm{G}}(\mathbf{x}) = \log \frac{p(y = 1)}{p(y = 2)} + \log \frac{p(\mathbf{x}^1|y = 1)}{p(\mathbf{x}^1|y = 2)} + \log \frac{p(\mathbf{x}^2|y = 1)}{p(\mathbf{x}^2|y = 2)} \, .$$

Such a representation can be obtained by assuming the generative DGP

$$p(\mathbf{x}|y) = w(\mathbf{x}^1, \mathbf{x}^2)p(\mathbf{x}^1|y)p(\mathbf{x}^2|y) \, ,$$

where $w(\mathbf{x}^1, \mathbf{x}^2)$ can be regarded as a normalisation factor. However, if, for all y, $p(\mathbf{x}^1|y)$ and $p(\mathbf{x}^2|y)$ are proper marginal distributions derived from $p(\mathbf{x}|y)$ (*i.e.*, $p(\mathbf{x}^1|y) = \sum_{\mathbf{x}^2} p(\mathbf{x}|y)$, $p(\mathbf{x}^2|y) = \sum_{\mathbf{x}^1} p(\mathbf{x}|y)$ and $\sum_{\mathbf{x}} p(\mathbf{x}|y) = \sum_{\mathbf{x}^1} p(\mathbf{x}^1|y) = \sum_{\mathbf{x}^2} p(\mathbf{x}^2|y) = 1$), then $w(\mathbf{x}^1, \mathbf{x}^2) \equiv 1$, given that there exists $\mathbf{x} = x$ such that $p(x|y = 1) \neq p(x|y = 2)$. In other words, it leads to assuming conditional independence between partial feature vectors $\mathbf{x}^1|y$ and $\mathbf{x}^2|y$ such that $p(\mathbf{x}|y) = p(\mathbf{x}^1|y)p(\mathbf{x}^2|y)$. In addition, to some extent, for a simple implementation in practice, Raina et al. (2003) further assume that $p(\mathbf{x}^1|y) = \prod_{j=1}^{h_1} p(x_j|y)$ and $p(\mathbf{x}^2|y) =$

$\prod_{j=h_1+1}^{h} p(x_j|y)$; these imply the conditional independence of the elements within $\mathbf{x}^1$ and $\mathbf{x}^2$ given y, respectively.

Raina et al. (2003) introduce two additional parameters θ_1 and θ_2 into the discriminant criterion, leading to different weights for different partial feature vectors in the discrimination. Two ways of weighting are proposed by Raina et al. (2003): one corresponds to assigning $\mathbf{x}$ to the group $y = 1$ if

$$p(y=1)p(\mathbf{x}^1|y=1)^{\frac{\theta_1}{h_1}}p(\mathbf{x}^2|y=1)^{\frac{\theta_2}{h_2}} \geq p(y=2)p(\mathbf{x}^1|y=2)^{\frac{\theta_1}{h_1}}p(\mathbf{x}^2|y=2)^{\frac{\theta_2}{h_2}} ,$$

which is the criterion (denoted by *Criterion-H*) corresponding to the normalised hybrid algorithm; the other gives

$$p(y=1)p(\mathbf{x}^1|y=1)^{\theta_1}p(\mathbf{x}^2|y=1)^{\theta_2} \geq p(y=2)p(\mathbf{x}^1|y=2)^{\theta_1}p(\mathbf{x}^2|y=2)^{\theta_2} ,$$

which is the criterion corresponding to the unnormalised hybrid algorithm. Without loss of generality, in this chapter we focus on the normalised hybrid algorithm.

Let us write $\theta = (\theta_1, \theta_2)^T$. Then the hybrid algorithm can be derived from

$$p_\theta(\mathbf{x}|y) = w_\theta(\mathbf{x}^1, \mathbf{x}^2)p(\mathbf{x}^1|y)^{\frac{\theta_1}{h_1}}p(\mathbf{x}^2|y)^{\frac{\theta_2}{h_2}} \text{ and } p_\theta(\mathbf{x}, y) = p(y)p_\theta(\mathbf{x}|y) ,$$

where $w_\theta(\mathbf{x}^1, \mathbf{x}^2)$ is independent of groups y so that it is cancelled out from *Criterion-H*, but it is not necessarily further factorised as $w_\theta(\mathbf{x}^1, \mathbf{x}^2) = w_\theta^1(\mathbf{x}^1)w_\theta^2(\mathbf{x}^2)$. However, in order to maintain $p_\theta(\mathbf{x}|y)$ as a proper probability distribution (so that *Criterion-H* is derived from a proper probabilistic model), with the marginal distributions $p(\mathbf{x}^1|y) = \sum_{\mathbf{x}^2} p_\theta(\mathbf{x}|y)$ and $p(\mathbf{x}^2|y) = \sum_{\mathbf{x}^1} p_\theta(\mathbf{x}|y)$, it is required that, for all y,

$$\sum_{\mathbf{x}^2} w_\theta(\mathbf{x}^1, \mathbf{x}^2)p(\mathbf{x}^2|y)^{\frac{\theta_2}{h_2}} = p(\mathbf{x}^1|y)^{1-\frac{\theta_1}{h_1}} ,$$

$$\sum_{\mathbf{x}^1} w_\theta(\mathbf{x}^1, \mathbf{x}^2)p(\mathbf{x}^1|y)^{\frac{\theta_1}{h_1}} = p(\mathbf{x}^2|y)^{1-\frac{\theta_2}{h_2}} .$$

In some cases, it might be difficult to validate the existence of such a $w_\theta(\mathbf{x}^1, \mathbf{x}^2)$, *e.g.*, when $\frac{\theta_1}{h_1} = 1$ while $\frac{\theta_2}{h_2} \neq 1$ or vice versa, as the sums, in terms of $\mathbf{x}$, on the left-hand sides of the above equations have to become independent of y. In other cases, further assumptions might be needed to guarantee the existence. We illustrate this by assuming that $w_\theta(\mathbf{x}^1, \mathbf{x}^2)$ can be further factorised in terms of $w_\theta(\mathbf{x}^1, \mathbf{x}^2) = w_\theta^1(\mathbf{x}^1)w_\theta^2(\mathbf{x}^2)$; in other words, we assume conditional independence between $\mathbf{x}^1|y$ and $\mathbf{x}^2|y$. It follows that

$$p_\theta(\mathbf{x}|y) = w_\theta^1(\mathbf{x}^1)p(\mathbf{x}^1|y)^{\frac{\theta_1}{h_1}}w_\theta^2(\mathbf{x}^2)p(\mathbf{x}^2|y)^{\frac{\theta_2}{h_2}} ,$$

which also leads to *Criterion-H*. One option for $w_\theta(\mathbf{x}^1, \mathbf{x}^2)$ is, for all y,

$$w_\theta^1(\mathbf{x}^1) = q(y)p(\mathbf{x}^1|y)^{1-\frac{\theta_1}{h_1}},\ w_\theta^2(\mathbf{x}^2) = \frac{1}{q(y)}p(\mathbf{x}^2|y)^{1-\frac{\theta_2}{h_2}}\ ,$$

where $q(y)$ is a non-zero function used to cancel out terms in y within $p(\mathbf{x}^1|y)^{1-\frac{\theta_1}{h_1}}$ and $p(\mathbf{x}^2|y)^{1-\frac{\theta_2}{h_2}}$. If such a $w_\theta(\mathbf{x}^1, \mathbf{x}^2)$ cannot be found, *Criterion-H* is not a Bayes discriminant criterion derived from a proper probabilistic model; nevertheless, in practice it can still be used as a criterion for discrimination, although in this case the hybrid algorithm is no longer a true Bayes classifier and, under a $0-1$ loss function, it cannot provide a minimum Bayes error.

Under *Criterion-H*, we classify $\mathbf{x}$ into $y = 1$ if $p_\theta(\mathbf{x}, y = 1) \geq p_\theta(\mathbf{x}, y = 2)$. Given $p_\theta(\mathbf{x}, y) \neq 0$, the discriminant function $\lambda_H(\mathbf{x})$ of the hybrid algorithm can be expressed in terms of weighted likelihood ratios as

$$\lambda_H(\mathbf{x}) = \log\frac{p(y=1)}{p(y=2)} + \frac{\theta_1}{h_1}\log\frac{p(\mathbf{x}^1|y=1)}{p(\mathbf{x}^1|y=2)} + \frac{\theta_2}{h_2}\log\frac{p(\mathbf{x}^2|y=1)}{p(\mathbf{x}^2|y=2)}\ .$$

Therefore, $\lambda_H(\mathbf{x})$ can be viewed as a "weighted" version of the discriminant function $\lambda_{\mathrm{G}}(\mathbf{x})$ of the generative classifier; however, as mentioned above, in theory the hybrid algorithm should satisfy some conditions about the marginal distributions in order to make the underlying model probabilistically valid. In addition, as with $\lambda_{\mathrm{G}}(\mathbf{x})$, most parameters, such as those for $p(\mathbf{x}^1|y)$ and $p(\mathbf{x}^2|y)$, in $\lambda_H(\mathbf{x})$ are learnt by using a generative approach; only a few parameters, such as the two weights θ_1 and θ_2, are then learnt by using a discriminative approach based on the learning results (about $p(\mathbf{x}^1|y)$ and $p(\mathbf{x}^2|y)$) from the generative approach. Therefore, the hybrid algorithm can be regarded as a generative classifier since it assumes the DGP $p(\mathbf{x}|y)$ and thus $p(\mathbf{x}, y)$.

With the assumption of conditional independence between $\mathbf{x}^1|y$ and $\mathbf{x}^2|y$, it follows that the two class-conditional probabilities, $p(\mathbf{x}|y)$ and $p_\theta(\mathbf{x}|y)$, are related by

$$p_\theta(\mathbf{x}|y) = p(\mathbf{x}|y)\left\{w_\theta(\mathbf{x}^1, \mathbf{x}^2)p(\mathbf{x}^1|y)^{\frac{\theta_1}{h_1}-1}p(\mathbf{x}^2|y)^{\frac{\theta_2}{h_2}-1}\right\}\ .$$

This indicates that, in practice, the hybrid algorithm assumes a scaled DGP $p_\theta(\mathbf{x}|y)$ which scales the generative DGP $p(\mathbf{x}|y)$ by a function not only of the group label y but also of the feature vector $\mathbf{x}$.

4.1.2 Class-posterior Probabilities

The second perspective for interpreting the hybrid algorithm is via its modelling of class-posterior probabilities:

$$p_\theta(y|\mathbf{x}) = \frac{p_\theta(\mathbf{x}, y)}{p_\theta(\mathbf{x})} = \frac{p(y)w_\theta(\mathbf{x}^1, \mathbf{x}^2)p(\mathbf{x}^1|y)^{\frac{\theta_1}{h_1}}p(\mathbf{x}^2|y)^{\frac{\theta_2}{h_2}}}{p_\theta(\mathbf{x})} = \frac{p(y)p(\mathbf{x}^1|y)^{\frac{\theta_1}{h_1}}p(\mathbf{x}^2|y)^{\frac{\theta_2}{h_2}}}{p_\theta(\mathbf{x})/w_\theta(\mathbf{x}^1, \mathbf{x}^2)} ,$$

where $p_\theta(\mathbf{x}) = \sum_y p_\theta(\mathbf{x}, y) = p_\theta(\mathbf{x}, y = 1) + p_\theta(\mathbf{x}, y = 2)$. According to Bayes' Theorem, the class-posterior probabilities in terms of the generative DGP $p(\mathbf{x}|y)$ are $p(y|\mathbf{x}) = p(y)p(\mathbf{x}|y)/p(\mathbf{x})$; it follows that

$$p_\theta(y|\mathbf{x}) = p(y|\mathbf{x}) \left\{ w_\theta(\mathbf{x}^1, \mathbf{x}^2)p(\mathbf{x}^1|y)^{\frac{\theta_1}{h_1}-1}p(\mathbf{x}^2|y)^{\frac{\theta_2}{h_2}-1} \frac{p(\mathbf{x})}{p_\theta(\mathbf{x})} \right\} .$$

This indicates that the normalised hybrid algorithm assumes scaled class-posterior probabilities $p_\theta(y|\mathbf{x})$ which scale the posterior probabilities $p(y|\mathbf{x})$ by a function not only of the feature vector $\mathbf{x}$ but also of the group label y.

4.1.3 Loss Functions

In order to find the best classifier, one of the optimal criteria is to minimize the so-called unconditional or total risk:

$$R(\hat{y}) = E_y \left[E_{\mathbf{x}|y} \left[L\left(y, \hat{y}(\mathbf{x})\right) \right] \right] = E_{\mathbf{x}} \left[E_{y|\mathbf{x}} \left[L\left(y, \hat{y}(\mathbf{x})\right) \right] \right] .$$

Such a criterion suffices to minimize the Bayes error, also called Bayes risk,

$$E_{y|\mathbf{x}} \left[L\left(y, \hat{y}(\mathbf{x})\right) \right] = \sum_{y=1}^{K} p(y|\mathbf{x}) L(y, \hat{y}(\mathbf{x})) .$$

A simple and widely used loss function is a $0-1$ loss such that $L(y, \hat{y}(\mathbf{x})) = 1$ if $\hat{y} \neq y$ and 0 otherwise. This leads to a Bayes classifier, $\hat{y}(\mathbf{x}) = \operatorname{argmax}_y p(y|\mathbf{x})$.

Since there are many loss functions that can lead to the normalised hybrid algorithm, here we only present one loss function, fixing $L(y, \hat{y}(\mathbf{x})) = 0$ if $\hat{y} = y$.

Proposition 4.1.1 *If the number of groups is $K \geq 2$, and it is assumed that, given y, $L(y, \hat{y}(\mathbf{x})) = L_y$ is independent of $\hat{y}(\mathbf{x})$ if $\hat{y} \neq y$, then the hybrid algorithm proposed in Raina et al. (2003) can be obtained through minimising the Bayes error with a loss function $L(y, \hat{y}(\mathbf{x}))$ such that $L(y, \hat{y}(\mathbf{x})) = L_y$ if $\hat{y} \neq y$ and 0 otherwise, where*

$$L_y = \frac{p(\mathbf{x}^1|y)^{\frac{\theta_1}{h_1}}p(\mathbf{x}^2|y)^{\frac{\theta_2}{h_2}}}{p(\mathbf{x}|y)} ,$$

in which h_1 and h_2 are the dimensions of $\mathbf{x}^1$ and $\mathbf{x}^2$, and $\mathbf{x} = (\mathbf{x}^{1T}, \mathbf{x}^{2T})^T$. A generalisation of such a loss function is $L_y = \frac{p_\theta(\mathbf{x}|y)}{p(\mathbf{x}|y)}$.

Proof The Bayes error for a classifier $\hat{y}(\mathbf{x})$ with such a loss function $L(y, \hat{y}(\mathbf{x}))$ is minimised by

$$\hat{y}(\mathbf{x}) = \underset{\hat{y}}{\operatorname{argmin}} \sum_{y \neq \hat{y}} p(y|\mathbf{x}) L_y = \underset{\hat{y}}{\operatorname{argmin}} \sum_{y \neq \hat{y}} p(\mathbf{x}^1|y)^{\frac{\theta_1}{h_1}} p(\mathbf{x}^2|y)^{\frac{\theta_2}{h_2}} p(y)$$

$$= \underset{\hat{y}}{\operatorname{argmin}} -p(\mathbf{x}^1|\hat{y})^{\frac{\theta_1}{h_1}} p(\mathbf{x}^2|\hat{y})^{\frac{\theta_2}{h_2}} p(\hat{y}) = \underset{y}{\operatorname{argmax}}\, p(\mathbf{x}^1|y)^{\frac{\theta_1}{h_1}} p(\mathbf{x}^2|y)^{\frac{\theta_2}{h_2}} p(y) \ ,$$

which is *Criterion-H*. The proof for the generalisation of L_y can be obtained similarly by replacing $p(\mathbf{x}^1|y)^{\frac{\theta_1}{h_1}} p(\mathbf{x}^2|y)^{\frac{\theta_2}{h_2}}$ with $p_\theta(\mathbf{x}|y)$. ∎

From Proposition 4.1.1, we observe that the loss from misclassification by the hybrid algorithm depends on the accuracy of the approximation of the true DGP $p(\mathbf{x}|y)$ by the assumed one, $p_\theta(\mathbf{x}|y)$ say. The closer $p_\theta(\mathbf{x}|y)$ is to $p(\mathbf{x}|y)$, the closer can $L(y, \hat{y}(\mathbf{x}))$ be approximated by a $0-1$ loss function. Furthermore, in contrast to the $0-1$ loss, L_y is dependent on $\mathbf{x}$.

4.1.4 A Multi-class Extension

Fujino et al. (2007) present the result of a multi-class and multi-partition extension of the hybrid algorithm by maximising a conditional entropy of $p(y|\mathbf{x})$ under certain constraints associated with joint distribution $p(\mathbf{x}, y)$ and class-conditional probabilities $p(\mathbf{x}^r|y)$ for each partial feature vector $\mathbf{x}^r, r = 1, \ldots, R$, as

$$p(y|\mathbf{x}) = \frac{e^{\mu_y} \prod_{r=1}^{R} p(\mathbf{x}^r|y)^{\lambda_r}}{\sum_y e^{\mu_y} \prod_{r=1}^{R} p(\mathbf{x}^r|y)^{\lambda_r}} \ ,$$

where λ_r and μ_y are Lagrange multipliers. This result is equivalent to a straightforward extension of the hybrid algorithm, in which $\lambda_r = \theta_r / h_r$ and $\mu_y = \log p(y) + \log w_\theta(\mathbf{x})$.

4.2 Parameter Estimation, Implementation and Evaluation of the Classifiers

4.2.1 Discriminative Learning of θ

By "hybrid", the normalised hybrid algorithm proposed in Raina et al. (2003) means to use a discriminative approach to the estimation of θ such that

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{m} \log p_\theta(y^{(i)}|\mathbf{x}^{(i)}) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{m} \log \frac{p_\theta(\mathbf{x}^{(i)}, y^{(i)})}{\sum_y p_\theta(\mathbf{x}^{(i)}, y)} ,$$

where m is the number of independent training observations $\left\{ \left(\mathbf{x}^{(i)}, y^{(i)}\right) \right\}_{i=1}^{m}$, in which $\left(\mathbf{x}^{(i)}\right)^T = \left(\left(\mathbf{x}^{1,(i)}\right)^T, \left(\mathbf{x}^{2,(i)}\right)^T \right)$. If y is a binary variable such that $y \in \{1, 2\}$, $p_\theta(y = 1|\mathbf{x})$ can be written in a way similar to that of logistic regression:

$$p_\theta(y = 1|\mathbf{x}) = \frac{\exp(\lambda_H(\mathbf{x}))}{1 + \exp(\lambda_H(\mathbf{x}))} ,$$

where $\lambda_H(\mathbf{x})$, as defined in Section 4.1.1, is the discriminant function corresponding to *Criterion-H*. As with linear logistic regression, $\lambda_H(\mathbf{x})$ is a linear function of θ_1 and θ_2.

Instead of using maximisation, we minimise the negative loglikelihood $-\ell_H$ to estimate θ_1 and θ_2, where

$$-\ell_H = -\sum_{i=1}^{m} \log p_\theta(y^{(i)}|\mathbf{x}^{(i)})$$

$$= \sum_{i=1}^{m} \left\{ \mathbf{1}_{\{y^{(i)}=1\}} \log \left(1 + e^{-\lambda_H(\mathbf{x}^{(i)})}\right) + \mathbf{1}_{\{y^{(i)}=2\}} \log \left(1 + e^{\lambda_H(\mathbf{x}^{(i)})}\right) \right\} .$$

Concerning $\lambda_H(\mathbf{x})$, in order to estimate the parameters in the same discriminative way as that of linear logistic regression, Raina et al. (2003) redefine θ as $\theta = (\theta_0, \theta_1, \theta_2)^T$, where $\theta_0 = \log \frac{p(y=1)}{p(y=2)}$, similar to the intercept in a linear logistic regression model, is estimated discriminatively, *i.e.*, $\log \frac{p(y=1)}{p(y=2)}$ is not calculated by using generative estimators of $p(y = 1)$ and $p(y = 2)$ but is directly estimated by a discriminative approach. Except for that, $\log \frac{p(\mathbf{x}^1|y=1)}{p(\mathbf{x}^1|y=2)}$ and $\log \frac{p(\mathbf{x}^2|y=1)}{p(\mathbf{x}^2|y=2)}$ are estimated by a generative approach.

Considering that the discriminative estimator of θ uses outputs from the generative estimator of $p(\mathbf{x}|y)$ as inputs while both estimators use the same training set, Raina et al. (2003) suggest that the discriminative estimator of θ is biased. Consequently, they use a "leave-one-out" strategy as follows:

$$\hat{\theta}_{-i} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{m} \log p_{\theta,-i}(y^{(i)}|\mathbf{x}^{(i)}) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{m} \log \frac{p_{\theta,-i}(\mathbf{x}^{(i)}, y^{(i)})}{\sum_y p_{\theta,-i}(\mathbf{x}^{(i)}, y)} ,$$

where $p_{\theta,-i}(\mathbf{x}^{(i)}, y)$ and $p_{\theta,-i}(\mathbf{x}^{(i)}, y^{(i)})$ are obtained from the data with the i-th observation removed. However, when the training set size m is large enough, there is little difference between $\hat{\theta}_{-i}$ and $\hat{\theta}$, and thus such a bias can be ignored. Therefore, in our study, we do not use the "leave-one-out" strategy to estimate θ.

4.2.2 Implementation of the Classifiers

In order to evaluate the discrimination performance of the hybrid algorithm, we compare it with two widely-used discriminative and generative classifiers, linear logistic regression and the naïve Bayes classifier, using simulated continuous and discrete data.

The naïve Bayes classifier is implemented by an R function *naiveBayes* from a contributed package **e1071** for R. As with Raina et al. (2003), for discrete data, we use Laplace (add-one) smoothing. For simulated continuous data, the naïve Bayes classifier, which assumes normal distributions for class-conditional probabilities $p(\mathbf{x}|y)$, corresponds to LDA-Λ when the covariance matrix Σ_1 of the group $y = 1$ is equal to the covariance matrix Σ_2 of the group $y = 2$, and corresponds to quadratic normal discriminant analysis with a common diagonal covariance matrix (QDA-Λ) when $\Sigma_1 \neq \Sigma_2$. The naïve Bayes classifier assumes the conditional independence of all h features given the group label y, such that $p(\mathbf{x}|y) = \prod_{j=1}^{h} p(x_j|y)$; its discriminant function $\lambda_{\mathrm{G}}(\mathbf{x})$ can be written as

$$\lambda_{\mathrm{G}}(\mathbf{x}) = \log \frac{p(y=1)}{p(y=2)} + \sum_{j=1}^{h} \log \frac{p(x_j|y=1)}{p(x_j|y=2)} \, .$$

The implementation of parameter estimation for the hybrid algorithm with $\lambda_H(\mathbf{x})$ consists of two steps: in the first step, by use of the R function *naiveBayes*, $p(x_j|y), j = 1, \ldots, h$, are generatively estimated and thus $\log \frac{p(\mathbf{x}^1|y=1)}{p(\mathbf{x}^1|y=2)}$ and $\log \frac{p(\mathbf{x}^2|y=1)}{p(\mathbf{x}^2|y=2)}$ can be calculated; in the second step, θ is estimated discriminatively by use of an R function *glm* (from a standard package **stats** in R) with $\log \frac{p(\mathbf{x}^1|y=1)}{p(\mathbf{x}^1|y=2)}$ and $\log \frac{p(\mathbf{x}^2|y=1)}{p(\mathbf{x}^2|y=2)}$ as predictor variables. The hybrid algorithm assumes conditional independence within the partial feature vectors, such that $p(\mathbf{x}^1|y) = \prod_{j=1}^{h_1} p(x_j|y)$ and $p(\mathbf{x}^2|y) = \prod_{j=h_1+1}^{h} p(x_j|y)$.

Linear logistic regression is implemented by the R function *glm* which uses an iteratively reweighted least squares algorithm (IRLS, or IWLS, also known as the Fisher scoring algorithm) to fit the model. The discriminant function $\lambda_{\mathrm{D}}(\mathbf{x})$ of linear logistic regression can be

written as

$$\lambda_{\mathrm{D}}(\mathbf{x}) = \beta_0 + \sum_{j=1}^{h} \beta_j x_j \ ,$$

which does not necessarily imply that the conditional independence assumption holds.

4.2.3 Evaluation of the Classifiers

To evaluate the performance of the three classifiers, we use the misclassification error rate (ER) and logarithmic loss (LL). The ER is defined as usual by the number of misclassified observations over the total number of observations; it is based on a $0-1$ loss function and is independent of the observed value x.

In contrast, the LL is dependent on x. The LL, also referred to as the logistic loss for logistic regression, is based on a loss function $L(y, \hat{y}(\mathbf{x})) = -\log p(y|\mathbf{x})$, where $p(y|\mathbf{x})$ is determined by the classifier $\hat{y}(\mathbf{x})$, and thus defined by

$$LL = \sum_{i=1}^{t} \left\{ -\log p(y^{(i)}|\mathbf{x}^{(i)}) \right\} \ ,$$

where t is the number of test observations. It can be easily recognised that the LL is in fact the negative of the log-likelihood of $p(y|\mathbf{x})$, and therefore the estimates obtained by the discriminative classifiers provide the best classification for the training observations if the minimum LL is used to measure the performance.

Consider two groups $y \in \{1, 2\}$ with the discriminant function $\lambda(\mathbf{x}) = \log \frac{p(y=1|\mathbf{x})}{p(y=2|\mathbf{x})}$. Then the LL can be rewritten as

$$LL = \sum_{i=1}^{t} \left\{ \left(-\log \frac{e^{\lambda(\mathbf{x}^{(i)})}}{1 + e^{\lambda(\mathbf{x}^{(i)})}} \right)^{\mathbf{1}_{\{y^{(i)}=1\}}} \left(-\log \frac{1}{1 + e^{\lambda(\mathbf{x}^{(i)})}} \right)^{\mathbf{1}_{\{y^{(i)}=2\}}} \right\} \ ,$$

where $\mathbf{1}_{\{y^{(i)}=k\}}$ is an indicator function of the subset $\{y^{(i)} = k\}$. A simple notation for the LL used by the machine learning community for two groups such that $y \in \{-1, 1\}$ is

$$LL = \sum_{i=1}^{t} \left\{ -\log \frac{1}{1 + e^{-y^{(i)}\lambda(\mathbf{x}^{(i)})}} \right\} = \sum_{i=1}^{t} \left\{ \log \left(1 + e^{-y^{(i)}\lambda(\mathbf{x}^{(i)})} \right) \right\} \ .$$

4.3 Numerical Studies

4.3.1 Simulation Studies

Twelve datasets are simulated here, of which 6 are composed of h continuous features and the other 6 are composed of h discrete features. In each continuous dataset, the data arise from

two h-variate normal distributions; in each discrete dataset, the data arise from two h-variate Bernoulli distributions.

Each dataset consists of $N = 10^3$ observations, which are equally categorised into two groups by a group label $y \in 1, 2$. Amongst them, $m/2$ observations from each of the two groups are used as training observations; m is sampled within $[100, 400]$ in steps of 25. For each sampled m, the N observations are randomly split into m training observations and $t = N - m$ test observations with 400 replicates; from them, the medians of the ERs and LLs are recorded and plotted. In each dataset, we set $h = 4$ and the feature vector $\mathbf{x} = (x_1, x_2, x_3, x_4)^T$ is composed of 2 partial feature vectors $\mathbf{x}^1 = (x_1, x_2)^T$ and $\mathbf{x}^2 = (x_3, x_4)^T$, *i.e.*, $h_1 = h_2 = 2$.

Amongst the 12 datasets, 6 datasets (3 continuous and 3 discrete) have $\Sigma_1 = \Sigma_2$, *i.e.*, the two groups have a common covariance matrix Σ. In addition, there are 4 datasets (2 continuous and 2 discrete) with diagonal covariance matrices, and thus for them the assumption of conditional independence of all h features of $\mathbf{x}$ given y underlying the naïve Bayes classifier is satisfied. There are also 4 datasets with block-diagonal covariance matrices of two blocks, where one block consists of the h_1 features of $\mathbf{x}^1$ and the other consists of the h_2 features of $\mathbf{x}^2$, and thus for them the assumption of conditional independence between $\mathbf{x}^1$ and $\mathbf{x}^2$ given y is satisfied. The other 4 datasets have full covariance matrices such that each of the h features of $\mathbf{x}$ given y is dependent on the others.

As our results for the simulated discrete data showed similar patterns to those for the simulated continuous data, only the latter are presented below. The former can be found in the appendices of this thesis.

4.3.1.1 Continuous Data with a Common Covariance Matrix Σ

The first 3 datasets contain simulated continuous data arising from two 4-variate normal distributions: $\mathbf{x} \sim \mathcal{N}(\mu_1, \Sigma_1)$ for the group with $y = 1$ and $\mathbf{x} \sim \mathcal{N}(\mu_2, \Sigma_2)$ for $y = 2$, with $\mu_1 = (1.5, 0, 0.5, 0)^T$, $\mu_2 = (-1.5, 0, -0.5, 0)^T$, $\Sigma_1 = \Sigma_2 = \Sigma$ and Σ is

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & c & 0 & 0 \\ c & 1 & 0 & 0 \\ 0 & 0 & 1 & c \\ 0 & 0 & c & 1 \end{bmatrix} \text{ or } \begin{bmatrix} 1 & c & c & c \\ c & 1 & c & c \\ c & c & 1 & c \\ c & c & c & 1 \end{bmatrix}$$

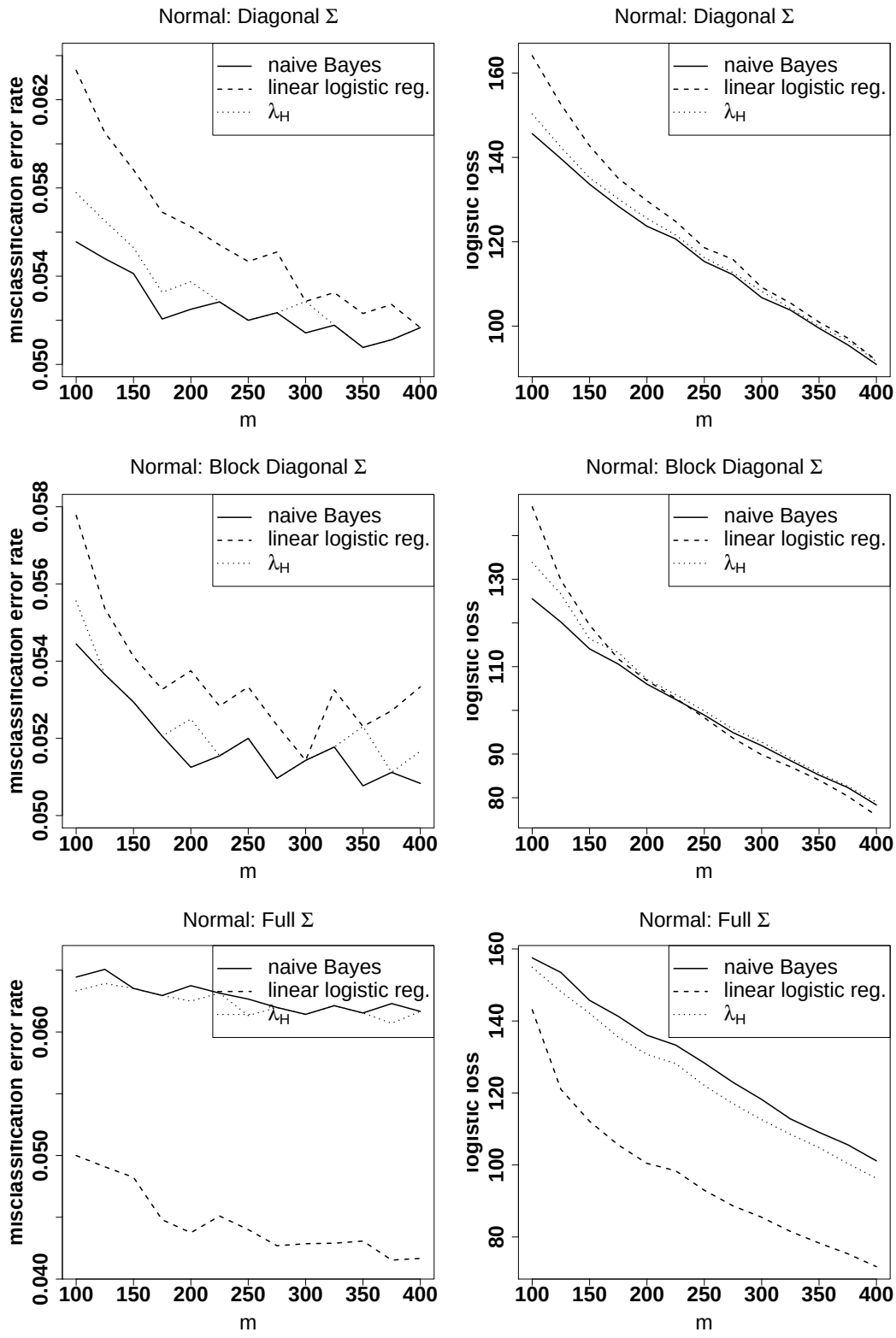


Figure 4.1: Simulated normally distributed data with equal covariance matrices. Plots of classification performance measured by ER and by LL vs. training set size m.

with $c = 0.25$, giving a diagonal, a block-diagonal and a full covariance matrix, respectively, for the 3 datasets.

Medians of the ERs and LLs are obtained from 400 replicates; the medians are plotted against the training set size m in Figure 4.1, of which each row represents the results for one dataset.

4.3.1.2 Continuous Data with Unequal Covariance Matrices Σ_1, Σ_2

The structure of the second set of 3 datasets is similar to that of the first set in Section 4.3.1.1, except that $\Sigma_1 \neq \Sigma_2$ and Σ_2 is

$$\begin{bmatrix} 0.25 & 0 & 0 & 0 \\ 0 & 0.75 & 0 & 0 \\ 0 & 0 & 1.25 & 0 \\ 0 & 0 & 0 & 1.75 \end{bmatrix}, \begin{bmatrix} 0.25 & c & 0 & 0 \\ c & 0.75 & 0 & 0 \\ 0 & 0 & 1.25 & c \\ 0 & 0 & c & 1.75 \end{bmatrix} \text{ or } \begin{bmatrix} 0.25 & c & c & c \\ c & 0.75 & c & c \\ c & c & 1.25 & c \\ c & c & c & 1.75 \end{bmatrix}$$

while Σ_1 is the same as Σ shown in Section 4.3.1.1, respectively for these 3 datasets. The results for these 3 datasets are shown in Figure 4.2.

4.3.2 Empirical Studies

For empirical studies, six continuous datasets in the UCI machine learning repository (Asuncion and Newman, 2007) are used here. The 6 UCI datasets are "Breast cancer Wisconsin (diagnostic)", "Breast cancer Wisconsin (prognostic)", "Connectionist bench (sonar)", "Ecoli (cp vs. pp)", "Pima Indians diabetes" and "Wine (1 vs. 2)".

Raina et al. (2003) used newsgroups data, reasonably dividing a message $\mathbf{x}$ into a message subject $\mathbf{x}^1$ and a message body $\mathbf{x}^2$ and obtaining very promising results from the hybrid algorithm. However, for these UCI datasets, there might not be such an apparently reasonable division. As a random division of $\mathbf{x}$ may break down the required connection of the features within either of the $\mathbf{x}^r$ and thus lead to a bias disfavouring the hybrid algorithm, we simply took the first half of the features as $\mathbf{x}^1$ and the others as $\mathbf{x}^2$. Such a simple division may preserve the connection between features, as similar features are in general next to each other in the order measured.

Similarly to the training-test split of the simulated datasets, for each group we randomly chose $\rho\%$ of the observations as training data and the remaining $(100-\rho)\%$ as test data, where

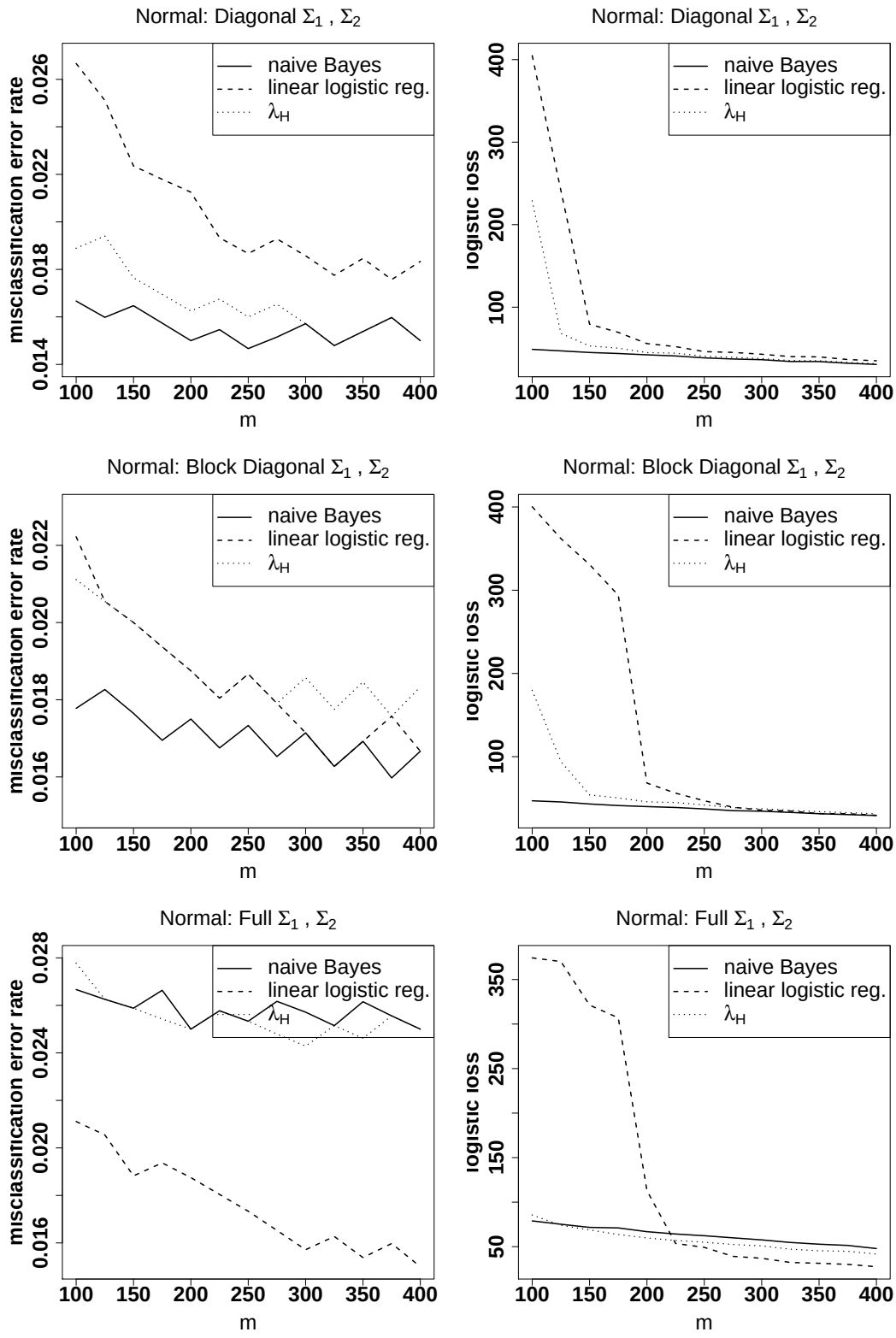


Figure 4.2: Simulated normally distributed data with unequal covariance matrices. Plots of classification performance measured by ER and by LL vs. training set size m.

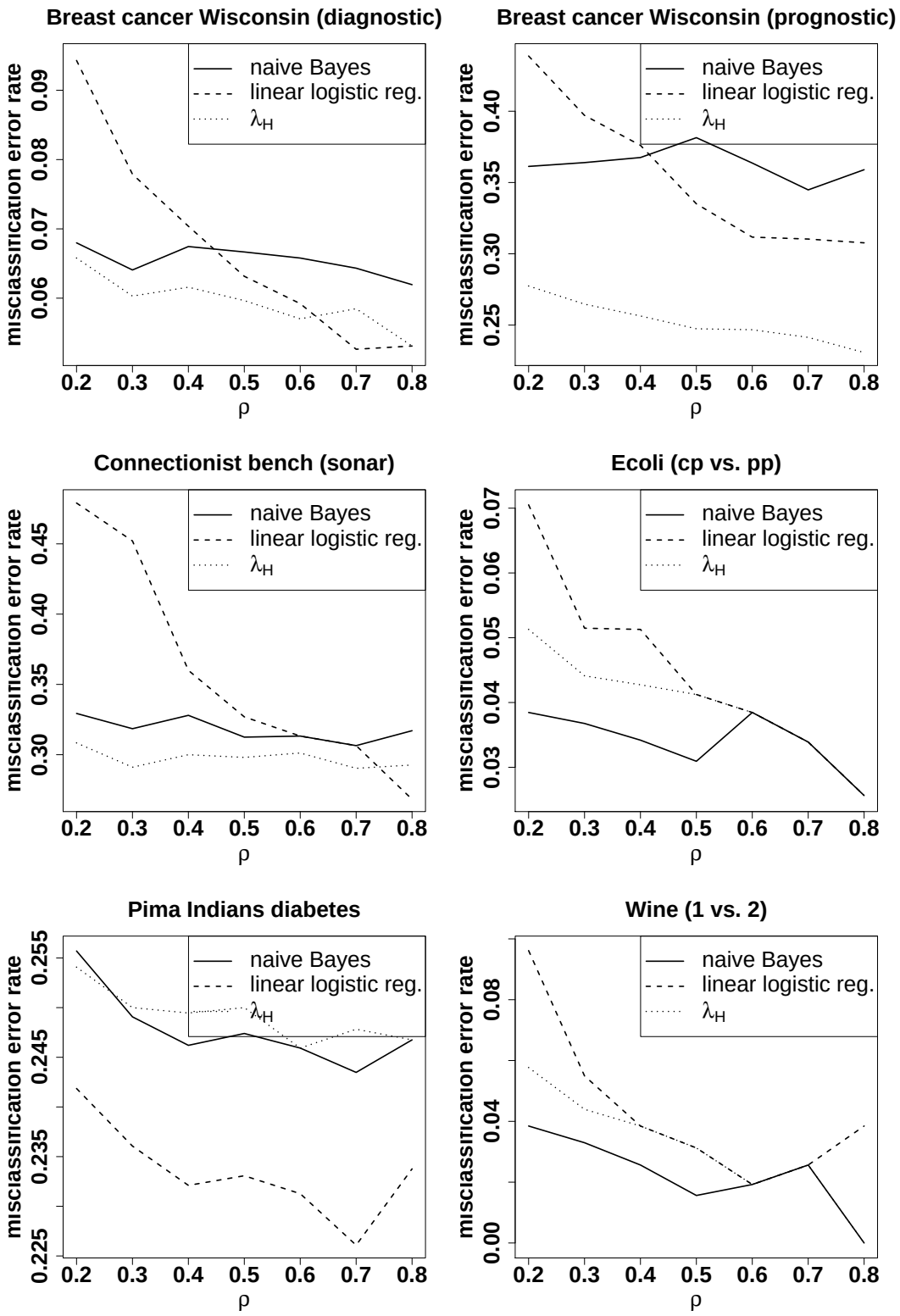


Figure 4.3: UCI datasets. Plots of classification performance measured by ER vs. ρ.

$\rho = 20(10)80$, such that the group proportion is preserved for training. For each value of ρ, we generated 100 such random partitions to assess classifier performance; medians of the ERs for these 100 replicates are shown in Figure 4.3, those of the LLs showing similar patterns.

4.3.3 Conclusions of Numerical Studies

Based on the results shown in Figure 4.1, 4.2 and 4.3, our numerical studies suggest the following conclusions.

First, with the simulated datasets, in general, in terms of both performance measures, namely ER and LL, if both the covariance matrices Σ_1 and Σ_2 are diagonal matrices, the naïve Bayes classifier performs the best; if both the covariance matrices Σ_1 and Σ_2 are full matrices, linear logistic regression performs the best, in particular when the training set size m is large. The superior performance of the naïve Bayes classifier can be attributed to the fact that the simulated data satisfy the assumption of conditional independence underlying the classifier; the superior performance of linear logistic regression can be attributed to its robustness when the assumptions underlying other classifiers are violated.

Secondly, the hybrid algorithm performs the best for 3 of the six UCI datasets while either the naïve Bayes classifier or linear logistic regression performs the best for the others.

Therefore, with these datasets, our studies suggest that the hybrid algorithm may provide worse performance than either the naïve Bayes classifier or linear logistic regression alone.

Chapter 5

Joint Generative-Discriminative Modelling Based on Statistical Tests for Classification

5.1 Introduction

The objective of statistical pattern classification is to classify an observation X into a group y, where X can be represented by a p-variate data vector $(x_1, \ldots, x_p)$ of its p measured variables and y is a categorical variable. The classification is based on a model, of which parameters are in general estimated from a training set of n labelled observations $\mathcal{X} = \{X_i = (x_{i1}, \ldots, x_{ip})\}_{i=1}^n$ with their labels $\mathcal{Y} = \{y_i\}_{i=1}^n$.

Based on our studies presented in Chapters 2, 3 and 4, in this chapter, we present a joint generative-discriminative modelling (JGD) approach to classification. This approach was also inspired by a suggestion, made but not developed in Rubinstein and Hastie (1997), that a promising hybrid approach is to 'partition the feature (variable) space into two. Train an informative model on those dimensions for which it seems correct, and a discriminative model on the others.' In other words, X is partitioned into two sub-vectors X_G and X_D, where $p(X_G|y)$ may be correctly modelled but $p(X_D|y)$ not, such that a generative approach is applied to X_G for $p(X_G|y)p(y)$ and a discriminative approach is applied to X_D for $p(y|X_D)$. Therefore, a key factor underlying the performance of such a classifier is the correctness of the partition of X, where confidence in $p(X_G|y)$ but not $p(X_D|y)$ should be based on the observed $\mathcal{X}$ and $\mathcal{Y}$.

The partition of variables into two subsets in our approach is based on statistical tests of the within-group distributions $p(x_p|y)$ of the variables x_p involved.

Closely-related work by Kang and Tian (2006) constructed an iterative partition of X, by starting with an empty $X_D^{(0)}$ (i.e., $X_G^{(0)} = X \setminus X_D^{(0)}$ is X), then, in the t-th iteration, moving from $X_G^{(t-1)}$ into $X_D^{(t-1)}$ a single variable x_j, namely the variable that can provide a classifier, which is based on $X_G^{(t)}$ and $X_G^{(t)}$, with the highest improvement of classification performance over the classifier that is based on $X_G^{(t-1)}$ and $X_G^{(t-1)}$; the procedure is continued till no such variable can be found. In each iteration, the classifier has to be applied $p_G^{(t-1)}$ times, where $p_G^{(t-1)}$ is the number of variables remaining in $X_G^{(t-1)}$, in order to select a within-loop winner.

In contrast to that of Kang and Tian (2006), the partition in our approach follows Rubinstein and Hastie (1997)'s suggestion that it should be based on different degrees of confidence we have in the distributions of $X_G|y$ and $X_D|y$. In addition, we do not partition variables in a heuristic or iterative way and thus only perform classification once rather than the many times (of the order of $p(p+1)$ times) necessary to compare the remaining variables in X_G. Therefore, our approach is much less intensive in computation, in particular for high-dimensional data.

We focus on two-group classification, in which y is a binary variable such that $y \in \{0, 1\}$ and the observations in the sample $\mathcal{X}$ are independent. The generalisation of our approach to multi-group scenarios is determined by the generalisation of corresponding generative and discriminative approaches involved.

5.2 Methodology

5.2.1 Models

A joint distribution $p(X, y)$ can be factorised into $p(X, y) = p(y|X)p(X)$, leading to discriminative approaches which assume the form of posterior probabilities $p(y|X)$ for classification, or into $p(X, y) = p(X|y)p(y)$, leading to generative approaches which assume a data-generating process (DGP) $p(X|y)$ for each group.

Suppose we know that, for the distribution $p(X_G|y)$, normality cannot be rejected, but, for $p(X_D|y)$, normality is rejected. Given $X = (X_D, X_G)$, it follows that there are several ways of factorising $p(X, y)$.

The first factorisation is

$$p(X, y) = p(X_D, X_G)p(y|X_D, X_G) \ , \tag{5.1}$$

which leads to a discriminative model for classification, which does not model the DGP $p(X_G|y)$, although we know that normality of $X_G|y$ cannot be rejected and therefore is plausible. One example of such a discriminative model is linear logistic regression (LLR).

The second factorisation is $p(X, y) = p(X_D, X_G|y)p(y)$, which gives

$$p(X, y) = p(y)p(X_G|y)p(X_D|X_G, y) \ , \tag{5.2}$$

the right-hand side of which includes a group distribution $p(y)$, a DGP $p(X_G|y)$ and a conditional DGP $p(X_D|X_G, y)$, leading to a generative model. The factor $p(y)$ can be assumed multinomial.

Based on different specifications for $p(X_G|y)$ and $p(X_D|X_G, y)$, many special cases can be derived of this generative model; one of them includes an assumption of conditional independence between X_D and X_G given y such that $p(X_D|X_G, y) = p(X_D|y)$. Equation (5.2) then simplifies to

$$p(X, y) = p(y)p(X_G|y)p(X_D|y) \ , \tag{5.3}$$

and can then lead to a block-wise generalisation of the naïve Bayes classifier (NBC); however, as either we know little about $p(X_D|y)$ or our hypothesis about the nature of $p(X_D|y)$, such as normality, is rejected, the NBC can be wrong in its model specification and thus the estimation of $p(X_D|y)$ is not correct, in particular for continuous X_D. This motivates the third factorisation of $p(X, y)$.

By exchanging X_D and y in (5.2), we obtain the third factorisation of $p(X, y)$ as

$$p(X, y) = p(X_D)p(y|X_D)p(X_G|X_D, y) \ , \tag{5.4}$$

the right-hand side of which includes a to-be-ignored distribution $p(X_D)$, a discriminative element $p(y|X_D)$ and a conditional DGP $p(X_G|X_D, y)$, leading to a joint generative-discriminative model. This model also includes many special cases, based on different specifications of $p(y|X_D)$ and $p(X_G|X_D, y)$. For example, if X_D is categorical, then both $p(y|X_D)$ and $p(X_G|X_D, y)$ can be accommodated by the NBC, or the former by logistic regression and the latter by the NBC.

5.2.2 Our JGD Approach

We focus on the scenario in which both X_D and X_G contain only continuous variables and the model is represented by equation (5.4). Although for such a scenario in theory we could assume that the distribution $p(X_G|X_D, y)$ is, for example, a Gaussian distribution, it is in practice hard to test this. For simplicity, we assume conditional independence such that $p(X_G|X_D, y) = p(X_G|y)$; this leads to the simplified version

$$p(X, y) = p(X_D)p(y|X_D)p(X_G|y) \ . \tag{5.5}$$

However, it can still be computationally expensive to test this assumption in practice for high-dimensional data in order to implement the partition of X into (X_D, X_G). Therefore, as usual, $p(X_G|y)$ is assumed to be normal, mainly for convenience, although it is still not easy to test such multivariate distributions.

The classification-related difference between equations (5.3) and (5.5) is equivalent to the difference between $p(y)p(X_D|y)$ and $p(y|X_D)$, which has been extensively studied before, mainly under the assumption that the model specification of $p(X_D|y)$ is correct. Here we concentrate on the case in which such a model specification, such as the normality of $p(X_D|y)$, has been rejected by statistical tests and thereby model mis-specification has occurred. In fact, this is why the partitioning of X into X_G and X_D is important.

In this context, our JGD approach can be described as follows.

First, we test the null hypothesis of normality of each variable x_j of X, and incorporate x_j in X_G if normality is not rejected at a prescribed significance level α and into X_D otherwise. Therefore, the partition of X into X_D and X_G is achieved by performing a univariate normality test p times. We use the univariate Shapiro-Wilk test for normality and set $\alpha = 0.01$. As α increases, the normality of more and more variables will be rejected and, consequently, the dimension of X_G will decrease. For low-dimensional data sets, such as some presented in Section 5.3, X_G may become empty with certain high values of α, such as 0.05 or higher. When either X_D or X_G turns out to be empty, the JGD approach degenerates to either a generative or a discriminative approach.

Secondly, when neither X_D nor X_G is empty, based on Bayes' theorem and equation (5.5), we use the following classification rule: a new observation $Z = (Z_D, Z_G)$ is classified into group $y = 1$ if

$$\log\left[\left\{\frac{p(y=1|Z_D)}{p(y=0|Z_D)}\right\}\left\{\frac{p(Z_G|y=1)}{p(Z_G|y=0)}\right\}\right] > 0 \ , \tag{5.6}$$

and $y = 0$ otherwise.

The left-hand side of equation (5.6) is the sum of two terms.

One is a discriminative term, $\log\{p(y = 1|Z_D)/p(y = 0|Z_D)\}$. It is the logit function of the posterior probability $p(y = 1|Z_D)$, and thus, if the LLR model is adopted, it can be represented by $\beta_0 + \sum_{j=1}^{p_D} \beta_j z_{Dj}$, where p_D is the dimension of Z_D, z_{Dj} are the variables in Z_D and β_j are the coefficients corresponding to z_{Dj}.

The other is a generative term, $\log\{p(Z_G|y = 1)/p(Z_G|y = 0)\}$. It is the log-likelihood ratio of Z_G between the two groups, and thus corresponds to normal-based linear/quadratic discriminant analysis (L/QDA) with equal/unequal covariance matrices across the two groups, given equal priors for the two groups.

If, as in Kang and Tian (2006), we further assume that the variables within X_G are conditionally independent, such that $p(X_G|y) = \prod_{j=1}^{p_G} p(x_{Gj}|y)$, where x_{Gj} are the variables in X_G and p_G is the dimension of X_G, then this generative term corresponds to L/QDA with equal/unequal diagonal covariance matrices, or the NBC. In other words, such an assumption justifies the use of the NBC for X_G. For high-dimensional data, such an assumption of independence may provide better classification results than using a full covariance structure (Bickel and Levina, 2004; Fan and Fan, 2007), with variable selection taken into account.

In this context, equation (4) can be re-written as

$$\beta_0 + \sum_{j=1}^{p_D} \beta_j z_{Dj} + \sum_{j=1}^{p_G} \left\{ \log \frac{\sigma_{Gj0}}{\sigma_{Gj1}} - \frac{(z_{Gj} - \mu_{Gj1})^2}{2\sigma_{Gj1}^2} + \frac{(z_{Gj} - \mu_{Gj0})^2}{2\sigma_{Gj0}^2} \right\} > 0 \,, \qquad (5.7)$$

where β_0 and β_j can be estimated by applying, for example, the method of iteratively reweighted least squares to the subset of $\mathcal{X}$ determined by X_D; $\mu_{Gj1}, \mu_{Gj0}, \sigma_{Gj1}$ and σ_{Gj0} are means and standard deviations of groups $y = 1$ and $y = 0$, respectively, and can be estimated by applying maximum likelihood estimation to the subset of $\mathcal{X}$ determined by X_G.

For high-dimensional data such that $p \gg n$, variable selection is commonly used before classification is performed (Fan and Fan, 2007; Hall et al., 2008). Variable selection can, on the one hand, make many traditional classification algorithms feasible, and, on the other hand, remove noisy, irrelevant variables and thus improve the classification performance.

If k variables with $k \leq n$ are selected, then classical methods such as the NBC and LLR, which were established for low-dimensional scenarios such that $p \leq n$, can be used effectively and this is also the case with our JGD approach.

5.3 Numerical Studies

5.3.1 UCI Data with $p \leq n$ and Gene Expression Data with $p \gg n$

We apply our JGD approach to 6 datasets with continuous variables in the UCI machine learning repository (Asuncion and Newman, 2007) and 3 gene-expression datasets. The 6 UCI datasets, satisfying $p \leq n$ ($p \leq 100$, $100 \leq n \leq 1000$), are "Breast cancer Wisconsin (diagnostic)", "Breast cancer Wisconsin (prognostic)", "Connectionist bench (sonar)", "Ecoli", "Haberman's survival" and "Wine".

The 3 gene-expression datasets are "Colon Cancer" (Alon et al., 1999), "Leukemia" (Golub et al., 1999) and "Prostate Cancer" (Singh et al., 2002). The Colon Cancer dataset consists of $p = 2000$ genes for $n = 62$ observations (40 tumour and 22 normal colon-tissue vectors). The Leukemia dataset consists of $p = 7129$ genes for $n = 72$ observations (47 acute lymphoblastic leukemia (ALL) and 25 acute myeloid leukemia (AML) data vectors). In the case of the Prostate Cancer dataset, there are $p = 12600$ genes for $n = 136$ observations (77 prostate tumours and 59 non-tumour prostate vectors).

For the gene-expression datasets, we first preprocess the data as did Dudoit et al. (2002), and then, based on training sets of observations, select k variables (genes) by using a tilting method proposed by Hall et al. (2008); k is set at 30, so that $k < n$. The preprocessing includes the following steps: truncating and censoring intensities to the interval $[100, 16000]$; removing genes which showed little variation in intensity across all the observations; transforming intensities to base-10 logarithms; and standardising each observation to have zero mean and unit variance.

Similarly to Kang and Tian (2006), in terms of misclassification error rate, we compare the JGD approach with the NBC, LLR and recursive partitioning and regression trees (rpart) methods. As Kang and Tian (2006) discretised all the continuous variables into ten equal-length intervals whereas we use continuous variable without discretisation, it may not be appropriate to compare our results with theirs. Nevertheless, our empirical and simulation studies, for low- or high-dimensional real and simulated data, can be regarded as a complement to their results on other UCI datasets.

The NBC and rpart methods are implemented by the R packages *e1071* and *rpart*, respectively; LLR is implemented by an R function *logitreg* (Venables and Ripley, 2002), using the BFGS algorithm.

Data	$n(n_0, n_1)$	$p(\tilde{p}_G)$	JGD	NBC	LLR	rpart	p-v (J-N)	p-v (J-L)	p-v (J-r)
Bcwd	569(357,212)	30(2)	0.035	0.070	0.035	0.088	0.008	1	0.016
Bcwp	194(148,46)	32(8)	0.264	0.300	0.308	0.325	0.016	0.445	0.203
Sonar	208(111,97)	60(2)	0.269	0.293	0.333	0.262	0.539	0.773	0.945
Ecoli	195(143,52)	5(2)	0	0.025	0.025	0.051	1	1	0.625
Haber	306(225,81)	3(1)	0.250	0.246	0.250	0.295	0.812	1	0.344
Wine	130(59,71)	13(7.5)	0	0	0.038	0.077	1	0.125	0.031
Colon	62(40,22)	30(20)	0.071	0.143	0.200	0.243	1	0.031	0.016
Leuke	72(47,25)	30(5)	0	0	0	0.134	1	1	0.031
Prost	136(59,77)	30(4)	0.077	0.154	0.154	0.113	0.473	0.094	1

Table 5.1: Description of the real datasets, medians of ER obtained from 10-fold cross-validation of our JGD approach, the NBC, LLR and rpart methods, and p-values for the Wilcoxon signed-rank test for pairs of our approach with each of the other classifiers. Notation: $n(n_0, n_1)$: the numbers of observations in the whole dataset, and for groups $y = 0$ and $y = 1$, respectively; p: the number of variables in X; $\tilde{p}_G$: the median number of variables in X_G; Bcwd: Breast cancer Wisconsin (diagnostic); Bcwp: Breast cancer Wisconsin (prognostic); Sonar: Connectionist bench (sonar); Ecoli: Ecoli (cp vs. pp); Haber: Haberman's survival; Wine: Wine (1 vs. 2); Colon: Colon Cancer; Leuke: Leukemia; Prost: Prostate Cancer.

The description of the datasets, medians of misclassification error rates (ER) obtained from 10-fold cross-validation of the compared classifiers and the p-values for the Wilcoxon signed-rank test for pairs made up of our approach with each of the other classifiers are listed in Table 5.1. For each fold of the 10-fold cross-validation, X_G and X_D can be different from those obtained in other folds, as can, for the high-dimensional gene-expression data, the selected k variables.

5.3.2 Simulated Data with Independent Normal and Gamma Distributions

As we know, two normally distributed groups of data can lead to a linear discriminant function if the two within-group covariance matrices Σ_1 and Σ_0, for groups $y = 1$ and $y = 0$ respectively, are equal, satisfying the assumption underlying LLR, and to a quadratic function otherwise. The normal-based NBC here assumes that $\Sigma_1 \neq \Sigma_0$ and thus assumes a quadratic discriminant function; however, it can provide a linear function for the case with $\Sigma_1 = \Sigma_0$, given that the estimated covariance matrices are approximately equal.

The Gamma distribution has, for $x \geq 0$, probability density function $G(x; \alpha, \eta) = x^{\alpha-1}\eta^{\alpha}e^{-\eta x}/\Gamma(\alpha)$, where the shape parameter $\alpha > 0$ and the inverse scale parameter (also called the rate) $\eta > 0$. It follows that, if variables in X_D are conditionally independent given y, a discriminative term can be derived from $\log\{p(y = 1|X_D)/p(y = 0|X_D)\}$ in the form

$$\log \frac{p(y = 1|X_D)}{p(y = 0|X_D)} = \beta_0 + \sum_{j=1}^{p_D} \beta_j x_{Dj} + \sum_{j=1}^{p_D} \gamma_j \log x_{Dj} \, , \tag{5.8}$$

where, with parameters for group y denoted by α_{jy} and η_{jy},

$$\beta_0 = \log \frac{p(y = 1)}{p(y = 0)} + \sum_{j=1}^{p_D} \left\{ \alpha_{j1} \log \eta_{j1} - \alpha_{j0} \log \eta_{j0} + \log \frac{\Gamma(\alpha_{j0})}{\Gamma(\alpha_{j1})} \right\} \, , \tag{5.9}$$

$$\beta_j = -(\eta_{j1} - \eta_{j0}) \, , \gamma_j = \alpha_{j1} - \alpha_{j0} \, . \tag{5.10}$$

Therefore, this represents a linear discriminative term that satisfies the assumption underlying LLR if $\alpha_{j1} = \alpha_{j0}$ and otherwise does not. In addition, it violates the assumption underlying the NBC which is based on normal distributions in our study.

To explore different scenarios involving satisfaction or violation of the underlying assumptions, we simulated 4 datasets, for combinations of normally distributed data (as X_G) with equal/unequal Σ_1 and Σ_0 and data (as X_D) from Gamma distributions with equal/unequal α_{j1} and α_{j0}, respectively.

Data	$X_G\|y=0, X_G\|y=1$	$X_D\|y=0, X_D\|y=1$	JGD	NBC	LLR
Sim1	$N(-1,9), N(1,9)$	$G(2,1/4), G(2,1/2)$	✓		✓
Sim2	$N(-1,9), N(1,9)$	$G(3,1/4), G(2,1/2)$			
Sim3	$N(-1,9), N(1,36)$	$G(2,1/4), G(2,1/2)$	✓		
Sim4	$N(-1,9), N(1,36)$	$G(3,1/4), G(2,1/2)$			

Table 5.2: Description of the simulated datasets. Notation: $N(\mu, \sigma^2)$; $G(\alpha, \eta)$; ✓ indicates cases in which the underlying assumptions are satisfied.

Data	JGD	NBC	LLR	p-v (J-N)	p-v (J-L)
Sim1	0.350	0.350	0.350	0.984	1
Sim2	0.225	0.200	0.200	0.562	0.250
Sim3	0.275	0.375	0.425	0.062	0.008
Sim4	0.200	0.250	0.250	0.438	0.375

Table 5.3: Medians of ER obtained from 10-fold cross-validation of our JGD approach, the NBC and LLR, and p-values for the Wilcoxon signed-rank test for pairs made up of our approach with each of the other classifiers.

For simplicity, for each simulated dataset, we set $p_G = p_D = 1$ and $n = 200$ with 100 observations from each group. The structure of the 4 datasets is shown in Table 5.2 and results about the corresponding ER obtained from 10-fold cross-validation are listed in Table 5.3. The specification of the class-conditional distributions in Table 5.2 is such that, within each simulation, the variances of the Gamma distributions closely match those of the normal distributions.

5.3.3 Summary of Numerical Studies

From the classification results shown in Tables 5.1 and 5.3, we observe the following.

First, our results for continuous UCI and gene-expression datasets demonstrate that the classification performance of the JGD approach is in general slightly superior to that of the NBC, LLR and rpart methods. Its lack of statistically significant superiority may be either due to imbalance between the numbers of variables of X_G and X_D or due to the small number of pairs (10 pairs from 10-fold cross-validation) in the Wilcoxon signed-rank tests.

Secondly, the results for "Sim2", "Sim4" and "Sim1" indicate that, when the underlying assumptions for each method are either violated or largely satisfied, the JGD, NBC and LLR approaches show similar performance.

Thirdly, the results for "Sim3" show that, when only its own underlying assumptions are satisfied, the JGD approach can perform significantly better than the NBC and LLR methods.

5.4 Conclusions

The JGD classification approach partitioned variables into two subsets based on statistical tests about within-group distributions of the variables, and then used generative approaches for the variables which passed the tests and discriminative approaches for the other variables. Such a statistical partition of variables and a probabilistic combination of generative and discriminative approaches led to promising classification performance of this approach for both low- and high-dimensional data, as demonstrated by our numerical studies for empirical and simulated data.

As explained at the end of Section 5.1, our approach is much more economical in terms of computation time than that by Kang and Tian (2006). We have concentrated on particular choices for the generative and discriminative components of our models, but the overall principle is quite general and can accommodate many other special versions. Of course, we must ensure that the assumptions underlying our generative components can be tested statistically.

Chapter 6

On Generative and Discriminative Hidden Markov Models

In this chapter, we study the assumption of "mutual information independence", which is used by Zhou (2005) for deriving an output-dependent hidden Markov model, the so-called discriminative HMM (D-HMM), in the context of determining a stochastic optimal sequence of hidden states. The assumption is extended to derive its generative counterpart, the G-HMM. In addition, state-dependent representations for two output-dependent HMMs, namely HMMSDO (Li, 2005) and D-HMM, are presented.

6.1 Introduction

Amongst the latent (hidden) variable models for structured data such as time series, hidden Markov models (HMMs) for discrete-valued hidden states and state-space models (SSMs) for continuous-valued hidden states are widely used.

Traditionally, an HMM is generative because it models a distribution $P(O_1^n|S_1^n)$, the data generation process (DGP) of the observed output sequence, $O_1^n = o_1, \ldots, o_n$, given the hidden state sequence, $S_1^n = s_1, \ldots, s_n$, and thus $P(O_1^n|S_1^n)$, a state-dependent term, is included in the criterion for determining a stochastic optimal sequence of hidden states. Recently, Zhou (2005) proposed a discriminative hidden Markov model (D-HMM), which includes output-dependent terms $P(s_t|O_1^n), t = 1, \ldots, n$, in the criterion, based on an assumption of "mutual information independence". Meanwhile, Li (2005) presented the so-called "hidden Markov models with

states depending on observations" (HMMSDO), which assume that the current state s_t depends not only on the last state s_{t-1} but also on the last output o_{t-1}, so that output-dependent terms $P(s_t|s_{t-1}, o_{t-1})$ are included in the criterion.

Both the D-HMM and HMMSDO show superior performance in determining the optimal state sequence for certain applications. Zhou (2005) shows that the D-HMM outperforms the corresponding generative hidden Markov model (G-HMM) for part-of-speech tagging and phrase chunking; Li (2005) shows that HMMSDO outperforms the standard HMM for prediction of protein secondary structures when the training set is large enough.

6.2 Generative HMM

Following the notation used by Zhou (2005), the definition of the optimal hidden state sequence S_1^n based on the observed output sequence O_1^n is that of the maximum a posteriori (MAP) estimator S^* of S_1^n:

$$S^* = \underset{S_1^n}{\operatorname{argmax}} \left\{ \log P(S_1^n|O_1^n) \right\} \ . \tag{6.1}$$

The G-HMM rewrites the criterion (6.1) through applying Bayes' theorem and ignoring the item determined purely by O_1^n as

$$S^* = \underset{S_1^n}{\operatorname{argmax}} \left\{ \log P(S_1^n) + \log P(O_1^n|S_1^n) \right\} \ ,$$

which is further factorised as

$$S^* = \underset{S_1^n}{\operatorname{argmax}} \left\{ \log P(S_1^n) + \log \left(P(o_1|S_1^n) \prod_{k=2}^{n} P(o_k|O_1^{k-1}, S_1^n) \right) \right\} \ .$$

In order to make this formulation tractable, an assumption that O_1^n is conditionally independent given S_1^n is in general introduced as, for all $k \in \{2, \ldots, n\}$,

$$P(o_k|O_1^{k-1}, S_1^n) = P(o_k|S_1^n) \ , \tag{6.2}$$

and thus, based on such a conditional independence assumption, the MAP estimator for the G-HMM is simplified to

$$S^* = \underset{S_1^n}{\operatorname{argmax}} \left\{ \log P(S_1^n) + \sum_{i=1}^{n} \log P(o_i|S_1^n) \right\} \ . \tag{6.3}$$

The G-HMM is regarded as being generative because it directly models the DGP $P(o_i|S_1^n)$ of the observed o_i from the hidden S_1^n.

In practice, as for the standard HMM, the assumption (6.2) is further simplified to

$$P(o_k|O_1^{k-1}, S_1^n) = P(o_k|S_1^n) = P(o_k|s_k) \ , \tag{6.4}$$

and thus the MAP estimator of the standard HMM is

$$S^* = \underset{S_1^n}{\operatorname{argmax}} \left\{ \log P(S_1^n) + \sum_{i=1}^{n} \log P(o_i|s_i) \right\} \ . \tag{6.5}$$

6.3 Discriminative HMM from Mutual Information Independence

The D-HMM rewrites the criterion (6.1) through applying Bayes' theorem, but not ignoring the item determined purely by O_1^n, as

$$S^* = \underset{S_1^n}{\operatorname{argmax}} \left\{ \log P(S_1^n) + \log \frac{P(S_1^n, O_1^n)}{P(S_1^n)P(O_1^n)} \right\} \ .$$

To make this formulation tractable, an assumption that the mutual information ($MI\left(S_1^n, O_1^n\right) = \log \frac{P(S_1^n, O_1^n)}{P(S_1^n)P(O_1^n)}$) between S_1^n and O_1^n is independent with respect to each hidden s_i was introduced by Zhou (2005) as

$$MI(S_1^n, O_1^n) = \sum_{i=1}^{n} MI(s_i, O_1^n) \ , \tag{6.6}$$

or, in more detail,

$$\log \frac{P(S_1^n, O_1^n)}{P(S_1^n)P(O_1^n)} = \sum_{i=1}^{n} \log \frac{P(s_i, O_1^n)}{P(s_i)P(O_1^n)} = \sum_{i=1}^{n} \log \frac{P(s_i|O_1^n)}{P(s_i)} \ . \tag{6.7}$$

Based on such a representation, the MAP estimator for the D-HMM is simplified as (Zhou, 2005)

$$S^* = \underset{S_1^n}{\operatorname{argmax}} \left\{ \log P(S_1^n) + \sum_{i=1}^{n} \log P(s_i|O_1^n) - \sum_{i=1}^{n} \log P(s_i) \right\} \ . \tag{6.8}$$

The D-HMM is regarded as being discriminative because the criterion (6.8) includes directly the discriminative process $P(s_i|O_1^n)$, representing an output-dependence of a hidden state s_i on all the observed outputs O_1^n.

We shall make four observations about the D-HMM.

First, it is noted that the criterion (6.8) is simultaneously to maximise the maximum posterior marginal (MPM) estimator $\sum_{i=1}^{n} \log P(s_i|O_1^n)$ of $\log P(S_1^n|O_1^n)$ and to maximise the distance between the state transition model $\log P(S_1^n)$ and its independence-based counterpart $\sum_{i=1}^{n} \log P(s_i)$.

Secondly, in order to satisfy the assumption (6.7) underlying the D-HMM, it is required that

$$\prod_{k=2}^{n} \frac{P(s_k|S_1^{k-1}, O_1^n)}{P(s_k|S_1^{k-1})} = \prod_{k=2}^{n} \frac{P(s_k|O_1^n)}{P(s_k)} \ .$$

Since this is valid for any value of s_k, it follows that, for all $k \in \{2, \ldots, n\}$,

$$\frac{P(s_k|S_1^{k-1}, O_1^n)}{P(s_k|S_1^{k-1})} = \frac{P(s_k|O_1^n)}{P(s_k)} \ . \tag{6.9}$$

Thirdly, the assumption (6.7) can be rewritten as

$$\log \frac{P(S_1^n, O_1^n)}{P(S_1^n)P(O_1^n)} = \sum_{i=1}^{n} \log \frac{P(s_i, O_1^n)}{P(s_i)P(O_1^n)} = \sum_{i=1}^{n} \log \frac{P(O_1^n|s_i)}{P(O_1^n)} \ . \tag{6.10}$$

Based on such a representation, the MAP estimator (6.8) for the D-HMM can be rewritten, with the term $\sum_{i=1}^{n} \log P(O_1^n)$ determined purely by O_1^n being ignored, as

$$S^* = \underset{S_1^n}{\operatorname{argmax}} \left\{ \log P(S_1^n) + \sum_{i=1}^{n} \log P(O_1^n|s_i) \right\} \ . \tag{6.11}$$

Therefore, the D-HMM can also be represented as being generative because the criterion (6.11) includes a generative-like process $P(O_1^n|s_i)$, representing a state-dependence of all the observed outputs O_1^n on a hidden state s_i.

Fourthly, it can be seen that, when the assumption (6.6) of mutual information independence develops from independence between pairs (s_i, O_1^n) into that between local pairs (s_i, o_i) such that $MI(S_1^n, O_1^n) = \sum_{i=1}^{n} MI(s_i, o_i)$, the criteria (6.11) and (6.8) degenerate into the criterion (6.5), indicating that the D-HMM degenerates into the standard HMM.

6.4 Generative HMM from Mutual Information Independence

Furthermore, similarly to the assumption (6.6) proposed by Zhou (2005), an assumption that mutual information between S_1^n and O_1^n is independent with respect to each observed o_i can be introduced here as

$$MI(S_1^n, O_1^n) = \sum_{i=1}^{n} MI(S_1^n, o_i) \ , \tag{6.12}$$

or, in more detail,

$$\log \frac{P(S_1^n, O_1^n)}{P(S_1^n)P(O_1^n)} = \sum_{i=1}^{n} \log \frac{P(S_1^n, o_i)}{P(S_1^n)P(o_i)} = \sum_{i=1}^{n} \log \frac{P(o_i|S_1^n)}{P(o_i)} \ . \tag{6.13}$$

Based on such a representation, we can obtain another generative model and its MAP estimator, with the term $\sum_{i=1}^{n} \log P(o_i)$ determined purely by O_1^n being ignored, as

$$S^* = \underset{S_1^n}{\operatorname{argmax}} \left\{ \log P(S_1^n) + \sum_{i=1}^{n} \log P(o_i|S_1^n) \right\} . \tag{6.14}$$

This estimator is in fact the estimator (6.3) of the G-HMM, *i.e.*, the G-HMM can be derived under the assumption (6.12), a type of mutual information independence.

Similarly, we shall make three observations about this G-HMM, which is derived from mutual information independence.

First, in order to satisfy the assumption (6.13) of the G-HMM, it is required that, for all $k \in \{2, \ldots, n\}$,

$$\frac{P(o_k|O_1^{k-1}, S_1^n)}{P(o_k|O_1^{k-1})} = \frac{P(o_k|S_1^n)}{P(o_k)} . \tag{6.15}$$

Therefore, under the MAP criterion (6.1), the conditions (6.15) and (6.2) have the same effect on determining the optimal hidden S_1^n.

Secondly, the assumption (6.13) can be rewritten as

$$\log \frac{P(S_1^n, O_1^n)}{P(S_1^n)P(O_1^n)} = \sum_{i=1}^{n} \log \frac{P(S_1^n, o_i)}{P(S_1^n)P(o_i)} = \sum_{i=1}^{n} \log \frac{P(S_1^n|o_i)}{P(S_1^n)} . \tag{6.16}$$

Based on such a representation, the MAP estimator (6.14) for the G-HMM can be rewritten, with the terms related to $\log P(S_1^n)$ being combined, as

$$S^* = \underset{S_1^n}{\operatorname{argmax}} \left\{ (1-n) \log P(S_1^n) + \sum_{i=1}^{n} \log P(S_1^n|o_i) \right\} . \tag{6.17}$$

Therefore, in this sense, the G-HMM can also be represented as being discriminative because the criterion (6.17) includes a discriminative-like process $P(S_1^n|o_i)$, representing an output-dependence of all the hidden states S_1^n on an observed output o_i.

Thirdly, it can be seen that, when the assumption (6.12) of mutual information independence develops from independence between pairs (S_1^n, o_i) into that between local pairs (s_i, o_i) such that $MI(S_1^n, O_1^n) = \sum_{i=1}^{n} MI(s_i, o_i)$, the criteria (6.17) and (6.14) degenerate into the criterion (6.5), indicating that the G-HMM degenerates into the standard HMM.

6.5 Equivalence between G-HMM and D-HMM

Once we assume a fully independent mutual information between any state-output combination (s_i, o_j) as

$$MI(S_1^n, O_1^n) = \sum_{i=1}^{n}\sum_{j=1}^{n} MI(s_i, o_j) \,, \tag{6.18}$$

or, in more detail,

$$\begin{aligned} \log \frac{P(S_1^n, O_1^n)}{P(S_1^n)P(O_1^n)} &= \sum_{i=1}^{n}\sum_{j=1}^{n} \log \frac{P(s_i, o_j)}{P(s_i)P(o_j)} \\ = \sum_{i=1}^{n}\sum_{j=1}^{n} \log \frac{P(o_j|s_i)}{P(o_j)} &= \sum_{i=1}^{n}\sum_{j=1}^{n} \log \frac{P(s_i|o_j)}{P(s_i)} \,, \end{aligned} \tag{6.19}$$

this assumption results in two criteria, one generative and the other discriminative, with the MAP estimators as

$$S^* = \underset{S_1^n}{\operatorname{argmax}} \{\log P(S_1^n) + \sum_{i=1}^{n}\sum_{j=1}^{n} \log P(o_j|s_i)\} \,, \tag{6.20}$$

$$S^* = \underset{S_1^n}{\operatorname{argmax}} \left\{ \log P(S_1^n) + \sum_{i=1}^{n}\sum_{j=1}^{n} \log P(s_i|o_j) - \sum_{i=1}^{n} \{n \log P(s_i)\} \right\} \,, \tag{6.21}$$

respectively. These two criteria are equivalent.

In the context of determining an optimal sequence of hidden states, apart from the equivalence above, up to now, we find two occurrences of equivalence between a discriminative representation of the MAP criterion and its generative counterpart: one is for the D-HMM between the criteria (6.8) and (6.11), the other is for the G-HMM between the criteria (6.17) and (6.14).

We shall further illustrate such equivalence with two simple but related HMMs: one is a generative-like state-dependent model, which assumes that the current output o_t depends not only on the current state s_t but also on the last state s_{t-1}; the other is a discriminative-like output-dependent model, the so-called HMMSDO (Li, 2005), which assumes that the current state s_t depends not only on the last state s_{t-1} but also on the last output o_{t-1}.

The joint distribution of the first generative-like state-dependent model is

$$P(S_1^n, O_1^n) = P(s_1)P(o_1|s_1) \prod_{i=2}^{n} P(s_i|s_{i-1})P(o_i|s_i, s_{i-1}) \,. \tag{6.22}$$

This distribution can be rewritten as

$$\begin{aligned} P(S_1^n, O_1^n) &= P(o_1, s_1) \prod_{i=2}^{n} P(s_i, o_i | s_{i-1}) \\ &= P(o_1) P(s_1|o_1) \prod_{i=2}^{n} P(o_i|s_{i-1}) P(s_i|s_{i-1}, o_i) \ , \end{aligned} \tag{6.23}$$

which leads to a discriminative-like output-dependent part $P(s_i|s_{i-1}, o_i)$ in the distribution. In fact, the only difference between the probabilistic directed acyclic graphs (DAGs) corresponding to the joint distributions (6.22) and (6.23) is that directions of edges from s_i to o_i are reversed.

Similarly, the joint distribution of the discriminative-like output-dependent HMMSDO, with $P(s_i|s_{i-1}, o_{i-1})$ included, is (Li, 2005)

$$P(S_1^n, O_1^n) = P(s_1) P(o_1|s_1) \prod_{i=2}^{n} P(s_i|s_{i-1}, o_{i-1}) P(o_i|s_i) \ . \tag{6.24}$$

This distribution can be rewritten as

$$\begin{aligned} P(S_1^n, O_1^n) &= P(s_1) P(o_n|s_n) \prod_{i=2}^{n} P(s_i, o_{i-1}|s_{i-1}) \\ &= P(s_1) P(o_n|s_n) \prod_{i=2}^{n} P(s_i|s_{i-1}) P(o_{i-1}|s_i, s_{i-1}) \ , \end{aligned} \tag{6.25}$$

which leads to a no-longer discriminative-like output-dependence in the distribution. In fact, the difference between the DAGs corresponding to the joint distributions (6.24) and (6.25) is only in that directions of edges from s_i to o_{i-1} are reversed. In practice, whether or not $P(o_{i-1}|s_i, s_{i-1})$ is reasonable needs to be justified, because it means that the current output depends on the next state.

6.6 Conclusions

We suggest that the mutual information assumption (6.12) results in the G-HMM, while another mutual information assumption (6.6) results in the D-HMM. However, in practice, whether or not the assumptions are reasonable and how the corresponding HMMs perform can be data-dependent; research efforts to explore an adaptive switching between or combination of these two models may be worthwhile. Meanwhile, we suggest that the so-called output-dependent HMMs could be represented in a state-dependent manner, and vice versa, essentially by application of Bayes' theorem.

Chapter 7

On Generative and Discriminative Image Thresholding

In this chapter, we present discriminative approaches to histogram-based image thresholding, in which the optimal threshold is derived from the maximum likelihood based on the conditional distribution $p(y|x)$ of y, the class indicator of a grey level x, given x. The discriminative approaches can be regarded as discriminative extensions of the traditional generative approaches to thresholding, such as Otsu's method and Kittler and Illingworth's minimum error thresholding (MET).

7.1 Introduction

Image thresholding is a simple and widely-used technique for segmentation, partitioning a grey-level image into segments corresponding to different classes (Sahoo et al., 1988; Pal and Pal, 1993; Sezgin and Sankur, 2004), given that the classes to some extent can be distinguished by their grey levels. Most thresholding approaches are proposed for two-class binarisation and are based on the grey-level histogram of an image (Sahoo et al., 1988; Sezgin and Sankur, 2004; Glasbey, 1993; Trier and Jain, 1995). Two of the most popular approaches are Otsu's method (Otsu, 1979) and Kittler and Illingworth's minimum error thresholding (MET) (Kittler and Illingworth, 1986).

Given an image of N pixels, Otsu's method selects the optimal threshold t^* as

$$t^* = \underset{t\in[0,T-1]}{\operatorname{argmin}} \ \sigma_w^2(t) = \pi_0(t)\sigma_0^2(t) + \pi_1(t)\sigma_1^2(t) \ ,$$

where $[0, T]$ is the range of grey level, and $\pi_0(t)$ and $\sigma_0(t)$ are respectively the proportion of and standard deviation within class $\mathcal{C}_0(t)$, where $\mathcal{C}_0(t)$ includes all the pixels with grey levels x less than t, i.e., $\mathcal{C}_0(t) = \{i : 0 \leq x_i \leq t, 1 \leq i \leq N\}$; $\pi_1(t)$, $\sigma_1(t)$ and $\mathcal{C}_1(t)$ are defined similarly for the remaining pixels, and thus $\sigma_w^2(t)$ is called within-class variance. The MET method selects t^* as

$$t^* = \underset{t \in [0, T-1]}{\operatorname{argmin}} \ \pi_0(t) \log \frac{\sigma_0(t)}{\pi_0(t)} + \pi_1(t) \log \frac{\sigma_1(t)}{\pi_1(t)} \ ,$$

where $\pi_y \neq 0, y = 0, 1$, and in practice σ_y is nonzero. Research efforts have been made to unify these two approaches (Kurita et al., 1992; Yan, 1996).

Kurita et al. (1992) show that Otsu's method is equivalent to maximisation of the log-likelihood based on the conditional distribution $p(x|y)$, where x is the grey level and $y \in \{0, 1\}$ is the class indicator corresponding to x, under the assumption that the grey level within each class (denoted by $x|y$) follows a normal distribution $\mathcal{N}(\mu_y, \sigma_y^2)$ and $\sigma_0^2 = \sigma_1^2$. Kurita et al. (1992) also show that MET is equivalent to maximisation of the log-likelihood based on the joint distribution $p(x, y)$, under the assumption that $x|y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ and $\sigma_0^2 \neq \sigma_1^2$. Since $p(x, y) = \pi_y p(x|y)$, where $\pi_y = p(y)$, Otsu's method is also equivalent to maximisation of the log-likelihood based on $p(x, y)$ with $\pi_0 = \pi_1 = 0.5$. In this sense, both Otsu's method and MET assume a data-generating process (DGP) $p(x, y)$; therefore, we call such approaches generative thresholding approaches. As with Fisher's linear discriminant, Otsu's original method does not assume normally distributed classes or that $\sigma_0^2 = \sigma_1^2$; therefore, hereafter we refer, as Otsu's method, to the generative method to which it is equivalent, shown in Kurita et al. (1992).

Since $p(x, y) = p(x)p(y|x) \propto p(y|x)$, the MET method is also equivalent to minimisation of the logistic loss, which is based on $-\log p(y|x)$. Meanwhile, under the assumption of normal distributions, both Otsu's method and MET are equivalent to minimisation of the expected misclassification error rate. In other words, both methods seek t^* such that $p(\mathcal{C}_1(t^*)|x = t^*) = p(\mathcal{C}_0(t^*)|x = t^*)$, leading to alternative iterative implementations by solving

$$\log\{p(\mathcal{C}_1(t)|x)/p(\mathcal{C}_0(t)|x)\} = 0$$

for x and then updating t, $p(\mathcal{C}_1(t)$ and $p(\mathcal{C}_0(t)$ in each iteration (Kittler and Illingworth, 1986; Gonzalez and Woods, 2002).

For both Otsu's method and MET, the grey-level histogram is assumed to be an empirical realisation of a two-component normal mixture. However, such an assumption often cannot

be guaranteed for real images, leading to a major potential risk of model mis-specification when generative thresholding is applied. In two-class discrimination, there are discriminative approaches which do not assume any DGP and which can be less sensitive to model mis-specification than are corresponding generative approaches (Rubinstein and Hastie, 1997; Ng and Jordan, 2001). Therefore, in this chapter, we present discriminative approaches to histogram-based image thresholding. The optimal threshold is derived from the maximum log-likelihood based on the conditional distribution $p(y|x)$. The discriminative approaches can be regarded as discriminative extensions of the traditional generative approaches to thresholding, such as Otsu's method and MET.

7.2 Discriminative Thresholding

For two-class discrimination, in terms of minimum misclassification error rate, an optimal discriminant criterion for classifying an observation x into class $\mathcal{C}_1$ with $y = 1$ (or $\mathcal{C}_0$ with $y = 0$) is a discriminant function $g(x, \alpha) = \log\{p(\mathcal{C}_1|x)/p(\mathcal{C}_0|x)\} > 0$ (or ≤ 0). For a pixel in grey-level images, x is in general its grey level as a scalar. The most widely used discriminant functions are a linear function $g(x, \alpha) = \beta_0 + \beta_1 x$, where $\alpha = (\beta_0, \beta_1)^T$, and a quadratic function $g(x, \alpha) = \beta_0 + \beta_1 x + \beta_2 x^2$, where $\alpha = (\beta_0, \beta_1, \beta_2)^T$.

The $g(x, \alpha)$ can be derived from a generative classifier, such as normal-based linear/quadratic discriminant analysis where $\mathcal{N}(\mu_y, \sigma_y^2)$ is assumed as the DGP for class y and where it is assumed that $\sigma_0^2 = \sigma_1^2$ for the linear case and $\sigma_0^2 \neq \sigma_1^2$ for the quadratic case. It can also be derived from a discriminative classifier, such as linear/quadratic logistic regression, in which no DGP is assumed.

Here we derive a discriminative thresholding approach from maximisation of the log-likelihood based on the conditional distribution $p(y|x)$, which can be represented as a function of $g(x, \alpha)$.

As $g(x, \alpha) = \log\{p(y = 1|x)/p(y = 0|x)\}$, after some algebra we obtain

$$p(y = 1|x) = e^{g(x,\alpha)} / \left(1 + e^{g(x,\alpha)}\right) \ , \ p(y = 0|x) = 1/\left(1 + e^{g(x,\alpha)}\right) \ .$$

It follows that, for an image of N pixels $\{(x_i, y_i)\}_{i=1}^N$, where x_i and y_i are the grey level and class indicator of the i-th pixel, the log-likelihood $\ell(\alpha)$ based on $p(y_i|x_i)$ is

$$\ell(\alpha) = \sum_{i=1}^{N} g(x_i, \alpha) y_i - \sum_{i=1}^{N} \log\left(1 + e^{g(x_i,\alpha)}\right) \ .$$

Let $h(x)$, $x = 0, \ldots, T$, denote the grey-level histogram constructed from the N pixels. For histogram-based thresholding, a threshold t partitions $h(x)$ into two sets of grey levels and thus partitions the image into two classes of pixels, denoted by $\mathcal{C}_0(t)$ and $\mathcal{C}_1(t)$, such that $y_i = 0$ if $x_i \leq t$ and $y_i = 1$ otherwise. As y_i changes with t, and the parameter α of $g(x, \alpha)$ is estimated from $\{(x_i, y_i)\}_{i=1}^N$ by maximisation of $\ell(\alpha)$, we write $g(x, \alpha)$ as $g(x, \alpha(t))$ and $\ell(\alpha)$ can be rewritten as

$$\ell(\alpha(t)) = \sum_{x=t+1}^{T} h(x) g(x, \alpha(t)) - \sum_{x=1}^{T} h(x) \log \left(1 + e^{g(x, \alpha(t))}\right) .$$

In this context, the optimal threshold t^* can be determined discriminatively as

$$t^* = \underset{t}{\operatorname{argmax}} \ \ell(\hat{\alpha}(t)) ,$$

where $\hat{\alpha}(t)$, estimated from $\mathcal{C}_0(t)$ and $\mathcal{C}_1(t)$, is the maximum-likelihood estimator of α for a threshold t. Estimation of $\alpha(t)$ proceeds similarly to that for logistic regression models, using $\mathcal{C}_0(t)$ and $\mathcal{C}_1(t)$ as the training set. As there is no convenient analytical solution for α, discriminative thresholding is of higher computational complexity than generative thresholding.

The multi-threshold extensions of the discriminative thresholding approaches can be obtained by using the log-likelihood for a multinomial logit model, which is the multi-class generalisation of logistic regression.

When the DGP is known, a generative approach is to be preferred in general. However, for real-world application, the DGP is always unknown, in which case a generative approach has to assume a specific DGP. For different assumptions of the DGP, a generative approach can have different variants. For example, variants of MET include those for Poisson (Pal and Bhandari, 1993), Rayleigh (Xue et al., 1999), Nakagami-Gamma, Weibull and log-normal distributions (Moser and Serpico, 2006).

In contrast to generative thresholding, a discriminative approach to thresholding assumes the discriminant function $g(x, \alpha)$ rather than the DGP, and this may lead to more robust performance against the model mis-specification. As parameter estimation within discriminative approaches is in general harder than that in generative approaches (Rubinstein and Hastie, 1997), the computational complexity of discriminative thresholding is in general higher than that of generative thresholding, as in our implementation below.

For illustration, we present two discriminative thresholding approaches, which have the same formula but different α for $g(x, \alpha)$ as those for Otsu's method and MET, respectively.

As Otsu's method corresponds to a linear discriminant function and MET corresponds to a quadratic, we define the discriminative Otsu method as

$$t^* = \underset{t}{\operatorname{argmax}}\ \ell(\hat{\alpha}(t)) \text{ with } g(x, \alpha(t)) = \beta_0(t) + \beta_1(t)x\ ,$$

and the discriminative MET as

$$t^* = \underset{t}{\operatorname{argmax}}\ \ell(\hat{\alpha}(t)) \text{ with } g(x, \alpha(t)) = \beta_0(t) + \beta_1(t)x + \beta_2(t)x^2\ .$$

7.3 Experiments with Discriminative Thresholding

In this section, we compare the performance of generative and discriminative versions of Otsu's method and MET. Comparison of approaches to image thresholding requires an appropriate evaluation method, and numerous methods have been developed based on various criteria (Sahoo et al., 1988; Sezgin and Sankur, 2004; Zhang, 1996; Zhang et al., 2007). Roughly speaking, supervised evaluation is subjective, requiring a pre-segmented image as ground-truth; unsupervised evaluation is objective but prefers an approach appropriate for the underlying evaluation criteria.

As with Kittler and Illingworth (1986) and Kurita et al. (1992), we compare the thresholding approaches by using histograms constructed from simulated data. The data for each class are simulated from normal, Poisson, log-normal and two-component normal mixture distributions. Normal distributions are, as used for Otsu's method and MET (Kurita et al., 1992), the most-commonly used distributions in image processing; Poisson distributions are justified based on a theory of image formation (Pal and Bhandari, 1993); log-normal distributions are used as heavy-tailed adaptions of Rayleigh distributions for the thresholding of synthetic aperture radar (SAR) amplitude images (Moser and Serpico, 2006); and, compared to normal, Poisson and log-normal distributions, a normal mixture can be a better approximation to the distribution of a class in the histogram.

Although, in our scenario, the underlying distributions for the simulated data are known, they are unknown for real images. Therefore, we do not compare discriminant thresholding approaches versus a generative thresholding approach developed for a specific distribution, such as MET for Poisson distributions in Pal and Bhandari (1993) or for log-normal distributions in Moser and Serpico (2006).

For Otsu's method and MET, normally distributed classes can satisfy the underlying assumptions, while neither Poisson nor log-normally distributed data satisfy the assumptions.

For discriminant thresholding, as normal distributions are exponential families in canonical form, they satisfy the linear or quadratic formulation of $g(x, \alpha(t))$. Although Poisson distributions are also exponential families in canonical form, because of the equivalence of mean and variance, they only satisfy the linear formulation of $g(x, \alpha(t))$. Log-normal distributions are exponential families but not in canonical form; hence, they and normal mixture distributions satisfy neither the linear nor the quadratic formulation of $g(x, \alpha(t))$.

For Otsu's method and MET, the estimator of the parameter $\theta = (\pi_y, \mu_y, \sigma_y^2)^T$ is the maximum-likelihood estimator based on $p(x, y)$, which can be calculated directly from the histogram as in Otsu (1979), Kittler and Illingworth (1986) and Kurita et al. (1992). The thresholds obtained are denoted by t_O and t_M, respectively.

For discriminant thresholding, as for logistic regression, the estimator of the parameter α is implemented by an R function *glm* (from a standard package **stats**), which uses an iteratively re-weighted least squares algorithm to fit the model. The thresholds obtained are denoted by d_O and d_M, respectively.

We make following comments about our implementation. First, in order to avoid $\sigma_y = 0$, which may cause failure of MET, we only search for thresholds within the $[1, 99]$ percentile range of histograms. Secondly, since grey levels are in range of $[0, T]$, we left-truncate and right-censor the simulated data into that range.

We simulate six datasets, each with $10,000$ pixels, and set $T = 255$ as for 8-bit grey-level images. The datasets for normal distributions are unbalanced in terms of class proportions, while others are balanced. The setting of our simulated data is as follows.

The two datasets for normal distributions are the same as those used by Kurita et al. (1992): one has $\pi_1 = 0.05$, $\mu_1 = 50$, $\mu_2 = 150$ and $\sigma_1 = \sigma_2 = 18$; the other has $\pi_1 = 0.25$, $\mu_1 = 38$, $\mu_2 = 121$, $\sigma_1 = 8$ and $\sigma_2 = 40$.

As a Poisson distribution can be well approximated by a normal distribution when its mean is larger, such as 10, as with Pal and Bhandari (1993), we simulate pixels with low grey levels. The dataset for Poisson distributions has $\mu_1 = 5$, $\mu_2 = 20$. As the mean is equal to the variance for Poisson distributions, the two classes have unequal variances.

The dataset for log-normal distributions has logarithms having $\mu_1 = 2$, $\mu_2 = 4$, $\sigma_1 = 1/2$ and $\sigma_2 = 1/4$.

One of the two datasets for normal mixture distributions has four components, two for each class with equal mixing weights. The two components $\mathcal{N}(\mu_{1,a}, \sigma_1^2)$ and $\mathcal{N}(\mu_{1,b}, \sigma_1^2)$ for the first mixture are specified with $\mu_{1,a} = 60$ and $\mu_{1,b} = 80$; and the two components $\mathcal{N}(\mu_{2,a}, \sigma_2^2)$ and $\mathcal{N}(\mu_{2,b}, \sigma_2^2)$ for the second mixture are specified with $\mu_{2,a} = 120$ and $\mu_{2,b} = 140$. In addition, $\sigma_1^2 = \sigma_2^2 = 10$, and hence the two classes have equal variances. The other dataset for normal mixture distributions is the same as the previous one but with $\sigma_1^2 = 5$ and $\sigma_2^2 = 15$, and hence the two classes have unequal variances.

The thresholding results for these six datasets are shown in Figure 7.1. We observe the following.

For the datasets from normal distributions, where the histograms are themselves normal mixtures, the discriminative Otsu method (d_O) gives almost the same results as MET (t_M), which is better than the Otsu's original method (t_O) (Kurita et al., 1992) and the discriminative MET (d_M). The same phenomenon appears for the Poisson dataset. For the other three datasets, all the four methods of study show the similar thresholds and thus comparable performance.

Note that, for all six datasets, although the discriminative MET does not provide satisfactory results, the discriminative Otsu method consistently provides relatively good performance, compared to the original methods. In terms of the level of computational complexity, that of the discriminative Otsu method, which corresponds to a linear discriminant function, is lower than that of the discriminative MET, which corresponds to a quadratic, whereas those of both discriminative approaches are higher than those of the original approaches in parameter estimation.

7.4 Conclusions

The discriminative approach to histogram-based image thresholding proposed in this chapter is based on maximum likelihood corresponding to the conditional distribution $p(y|x)$, rather than $p(x, y)$ as in the case of the traditional generative thresholding. For our simulated datasets, results show that the discriminative Otsu method consistently provides relatively good performance. Considering its robustness and model simplicity, we suggest the use of the discriminative Otsu method for scenarios in which Otsu's original method and MET do not perform well due to model mis-specification and in which the computation is not demanding.

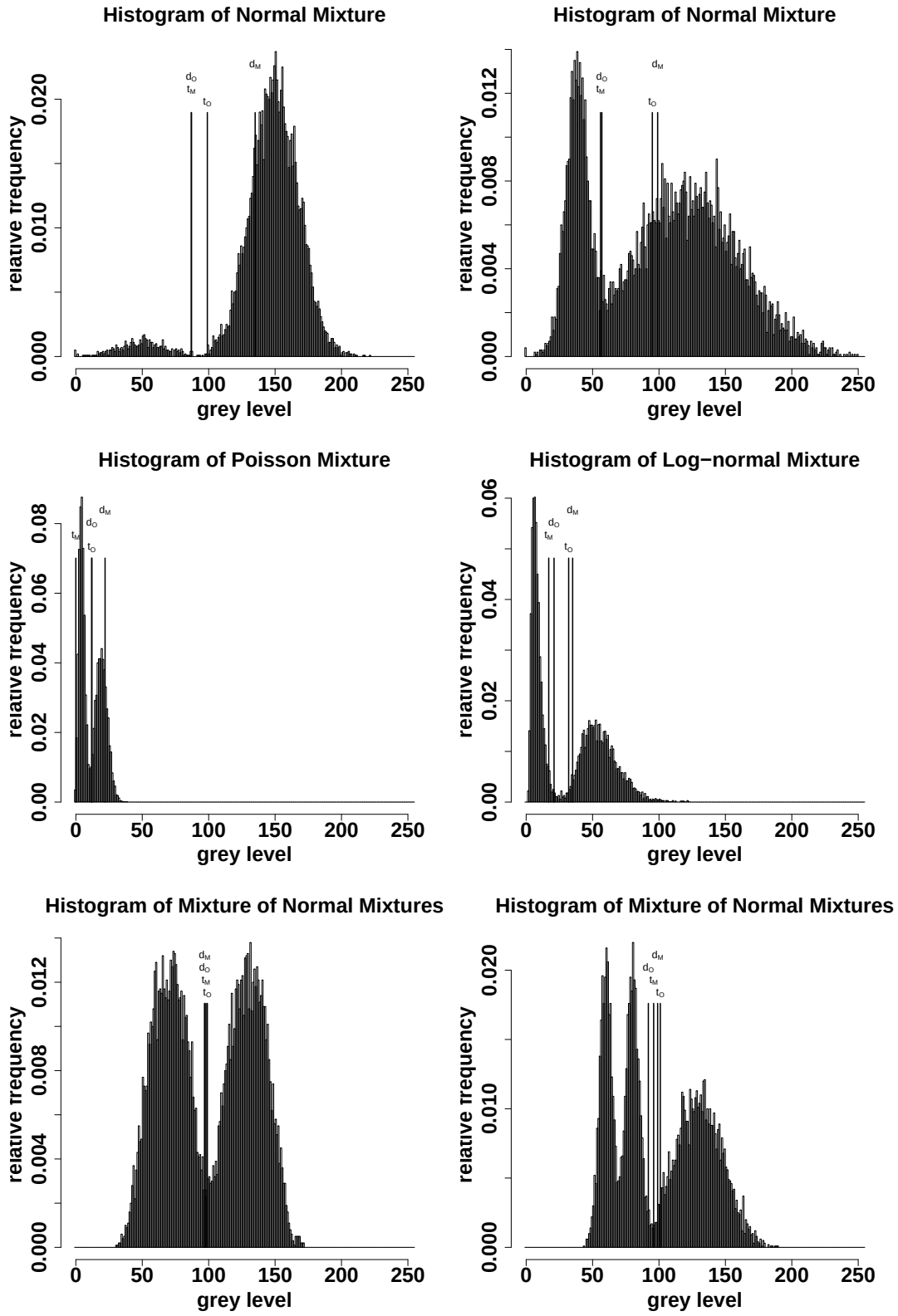


Figure 7.1: Thresholding results for 6 simulated datasets. Here t_O, t_M, d_O and d_M are thresholds from Otsu's method, MET and their discriminative counterparts, respectively.

Chapter 8

Summary, Conclusions, Discussion and Future Work

8.1 Summary of the Thesis

Classification is a ubiquitous problem tackled in statistics, machine learning, pattern recognition and data mining (Hand, 2006). The sampling and diagnostic paradigms for classification (Dawid, 1976; Titterington et al., 1981; Hand and Yu, 2001), studied before in the statistics community both theoretically and empirically, re-emerged in the machine learning community under the new terminology of generative and discriminative classifiers (Ng and Jordan, 2001), in particular with some hybrid modelling and learning techniques (Raina et al., 2003; Bouchard and Triggs, 2004; McCallum et al., 2006; Bishop and Lasserre, 2007) to exploit the best of both paradigms.

The purpose of this thesis was to investigate the degree of innovation and performance improvement made with these hybrid classifiers, and in the end, based on the investigation, to develop our own philosophy and techniques for classification.

The main approach used in the thesis towards its goal was to consider the hybrid classifiers together with some widely-used statistical classifiers, figuring out the underlying statistical assumptions and the connections between them, implementing simulation or empirical studies for them and comparing the corresponding results thereby obtained.

In Chapter 2, we performed some empirical and simulation studies to provide extension of and make comments on a highly-cited report (Ng and Jordan, 2001) which compared the naïve

Bayes classifier (NBC) or normal-based linear discriminant analysis (LDA) with linear logistic regression (LLR) and claimed that there exist two distinct regimes of performance between the generative and discriminative classifiers, depending on the training-set size m. However, our studies suggested that it is not so reliable to claim existence of the two distinct regimes and that pairing of either LDA assuming a common diagonal covariance matrix (LDA-Λ) or the NBC and LLR may not be perfect. Hence, it may not be reliable for any claim that was derived from the comparison between LDA-Λ or the NBC and LLR to be generalised to all generative and discriminative classifiers.

In Chapters 3 and 4, we studied extensively two hybrid-learning techniques, namely the hybrid generative-discriminative algorithm (Raina et al., 2003) and the generative-discriminative tradeoff (GDT) approach (Bouchard and Triggs, 2004). We argued that both the GDT and the hybrid algorithm are by nature generative models integrating both discriminative and generative learning. They are therefore still sensitive to model mis-specification of the data-generating process (DGP).

8.2 Conclusions

Based on the results from above investigations, our conclusions were as follows.

First, there was no universal winner amongst the generative, discriminative and hybrid classifiers; the performance is data-dependent, as shown in Chapters 2, 3 and 4.

This led to our second argument: it was recommended to first explore the data in order to validate the assumptions underlying candidate classifiers and then to decide to use either generative, discriminative or hybrid classifiers.

We developed such an argument by proposing, in Chapter 5, a joint generative-discriminative modelling (JGD) approach to classification, by partitioning variables into two subsets based on statistical tests of the DGP. Our JGD approach adopts statistical tests, such as normality tests, of the assumed DGP for each variable to justify the use of generative classifiers for the variables which satisfy the tests and of discriminative classifiers for the other variables. Such a partition of variables and a combination of generative and discriminative classifiers were derived in a probabilistic rather than a heuristic way, and also demonstrated promising performance for practical application to both low- and high-dimensional data.

Our third conclusion was that, considering the pairing of generative and discriminative

models, we could develop a discriminative counterpart for an existing generative approach and vice versa, as shown in Chapters 6 and 7. However, within such a pair, two models have in general different underlying assumptions, explicitly or implicitly; therefore, whether or not the assumptions are reasonable and how the corresponding pairs perform are again data-dependent.

8.3 Some Further Discussion

First, as discussed in Chapter 2, Ng and Jordan (2001) claimed that there exist two distinct regimes of performance between the generative and discriminative classifiers with regard to the training-set size m. They came to that conclusion by comparing the normal-based NBC and LLR, of which the NBC performs better with smaller m and LLR with larger m. A similar pattern of two distinct regimes with regard to m was also reported by Perlich et al. (2003), based on the performance of logistic regression (LR) and tree induction; they found that LR performs better with smaller m and tree induction with larger m. Therefore, although tree induction and LR are not a pair of generative and discriminative classifiers, it could be interesting to explore such a pattern for other pairs of classifiers.

Secondly, one of the key points of the hybrid algorithm in Raina et al. (2003) is to assign weights to the class-conditional distributions of subsets of variables $\mathbf{x}$; the subsets were obtained by partitioning $\mathbf{x}$. The extremes of such a block-wise NBC are either the independence model investigated by Titterington et al. (1981) and Hand and Yu (2001), assigning a common weight, or a more sophisticated model, assigning different weights to the distributions of different variables. In addition, it may not be necessary to use a hybrid strategy to estimate parameters, as the weights can be also estimated in a generative way.

Thirdly, although the hybrid classifiers, such as the GDT and the hybrid algorithm, offered good empirical results, our results showed that simpler generative classifiers like NBC and discriminative classifiers like LLR could offer comparable performance to the hybrid classifiers. This conformed to an argument made by Hand (2006) that simple classifiers typically yield performance that is almost as good as more sophisticated classifiers. Meanwhile, a generally-valid empirical evaluation of classifiers is always an important but difficult problem (Hand, 2006). Our setting of simulation and empirical studies in general followed or extended those of the original papers, such as in Chapters 2, 3 and 7, if practically possible. However, a more comprehensive comparative study may benefit from the theory of experimental design, after

investigation of the underlying assumptions of the classifiers under study.

Finally, some good performance of hybrid classifiers, such as the hybrid algorithm (Raina et al., 2003) and the NBC-based independence model (Titterington et al., 1981; Hand and Yu, 2001), may be the consequence of bias-variance trade-off, as they are in general biased models.

8.4 Potential Future Work

Based on the results presented in this thesis, several directions for future work merit investigation.

First, we could use resampling methods for high-dimensional low-sample-size data, such as bagging or boosting of simple classifiers like NBC which has shown good performance for high-dimensional data (Hand and Yu, 2001).

Secondly, we could compare generative and discriminative models for problems where the distribution of training samples is different from that of test samples, and then develop a hybrid classifiers for such a scenario.

Thirdly, one well-studied model corresponding to equation (5.2) is the general location model for mixed categorical and continuous data (Krzanowski, 1983), in which X_G contains categorical variables and X_D contains continuous variables. For such a model, corresponding to traditional generative approaches to its parameter estimation based in general on normal distributions, we could develop and validate discriminative modelling and learning approaches.

Finally, when causal, effect and background variables are candidate predictors for an outcome of response, such as in medical statistics with symptomatological, aetiological and patient's background variables, it could be better to select only causal variables as the predictors, as suggested by Ni Bhrolchain (1979). Approaches to achieving this may include weighting each variable (or their class-conditional distributions), or doing causality-based variable selection beforehand, although the latter could be a challenging task, which is beyond the topic of this thesis.

Appendix A

Appendix for Chapter 3

A.1 Asymptotic Efficiency of GDT for Linear Normal Discrimination

A.1.1 Linear Normal Discrimination

We assume that, within each sub-population, the feature vector $\mathbf{x}$ arises from one of two multivariate normal distributions with different means but the same covariance matrix, *i.e.*, $\mathbf{x}|\theta_1 \sim \mathcal{N}(\mu_1, \Sigma)$, $\mathbf{x}|\theta_0 \sim \mathcal{N}(\mu_0, \Sigma)$, and that no mis-specification occurs. In this context, a linear discriminant function is derived, as in Section 1.1.4:

$$g(\mathbf{x}, \alpha) = \log \frac{\pi_1}{\pi_0} - \frac{1}{2}(\mu_1^T W \mu_1 - \mu_0^T W \mu_0) + (\mu_1 - \mu_0)^T W \mathbf{x} = \beta_0 + \beta^T \mathbf{x} \,,$$

where $W = \Sigma^{-1}$, so that $\alpha^T = (\beta_0, \beta^T)$, $\theta^T = (\pi_1, \mu_1^T, \mu_0^T, (\text{vech}(W))^T)$.

A.1.2 Estimation of $\Sigma_{\text{g}}(\hat{\theta})$

Asymptotic properties of maximum likelihood estimators suggest that $\sqrt{n}(\hat{\theta} - \theta) \sim \mathcal{AN}(\mathbf{0}, \Sigma_{\text{g}}(\hat{\theta}) = nI_{\text{g}}^{-1}(\theta))$, where $I_{\text{g}}(\theta)$ is the Fisher information matrix,

$$I_{\text{g}}(\theta) = E\left\{ \frac{\partial \ell_{\text{g}}(\theta)}{\partial \theta} \frac{\partial \ell_{\text{g}}(\theta)}{\partial \theta^T} \right\} = E\left\{ -\frac{\partial^2 \ell_{\text{g}}(\theta)}{\partial \theta \partial \theta^T} \right\} \,.$$

After some algebra, we can obtain the following results:

$\sqrt{n}(\hat{\pi}_1 - \pi_1) \sim \mathcal{AN}(0, \pi_1 \pi_0)$,

$\sqrt{n}(\hat{\mu}_1 - \mu_1) \sim \mathcal{AN}(\mathbf{0}, \frac{1}{\pi_1}\Sigma)$, $\sqrt{n}(\hat{\mu}_0 - \mu_0) \sim \mathcal{AN}(\mathbf{0}, \frac{1}{\pi_0}\Sigma)$,

$\sqrt{n}(\text{vech}(\hat{W}) - \text{vech}(W)) \sim \mathcal{AN}(\mathbf{0}, nI_{\text{g}}^{-1}(\text{vech}(W)))$, where

$$\left\{I_{\text{g}}(\text{vech}(W))\right\}_{W_{i,j},W_{k,l}} = E\left\{-\frac{\partial^2 \ell_{\text{g}}(\theta)}{\partial W_{i,j}\partial W_{k,l}}\right\} = \frac{n(\Sigma_{i,k}\Sigma_{l,j} + \Sigma_{i,l}\Sigma_{k,j})}{(1+\delta_{i,j})(1+\delta_{k,l})} ,$$

in which $\Sigma_{i,j}$ and $W_{i,j}$ are the (i,j)-th components of Σ and W, respectively.

It follows that $\Sigma_{\text{g}}(\hat{\theta})$ is a block-diagonal matrix composed of a scalar $\Sigma_{\text{g}}(\hat{\pi}_1) = \pi_1\pi_0$, two $p \times p$ matrices $\Sigma_{\text{g}}(\hat{\mu}_1) = \frac{1}{\pi_1}\Sigma$ and $\Sigma_{\text{g}}(\hat{\mu}_0) = \frac{1}{\pi_0}\Sigma$, and a $\frac{p(p+1)}{2} \times \frac{p(p+1)}{2}$ matrix $\Sigma_{\text{g}}(\text{vech}(\hat{W})) = nI_{\text{g}}^{-1}(\text{vech}(W))$.

A.1.3 Estimation of $\Sigma_\lambda(\hat{\theta})$

Asymptotic properties of maximum likelihood estimators suggest that

$$\sqrt{n}(\hat{\theta} - \theta) \simeq \sqrt{n}\left[E\left\{-\frac{\partial^2 \ell_\lambda(\theta)}{\partial\theta\partial\theta^T}\right\}\right]^{-1} \cdot \frac{\partial \ell_\lambda(\theta)}{\partial\theta} \sim \mathcal{AN}(\mathbf{0}, \Sigma_\lambda(\hat{\theta})) ,$$

where $\ell_\lambda(\theta) = \lambda\ell_{\text{g}}(\theta) + (1-\lambda)\ell_{y|\mathbf{x}}(\theta)$, and $\Sigma_\lambda(\hat{\theta}) = nI_\lambda^{-1}(\theta)V_\lambda(\theta)I_\lambda^{-1}(\theta)$, in which, since $E\left\{\frac{\partial \ell_\lambda(\theta)}{\partial\theta}\right\} = 0$ and $\ell_{\text{g}}(\theta) = \ell_{y|\mathbf{x}}(\theta) + \ell_{\mathbf{x}}(\theta)$,

$$I_\lambda(\theta) = E\left\{-\frac{\partial^2 \ell_\lambda(\theta)}{\partial\theta\partial\theta^T}\right\} = \lambda I_{\text{g}}(\theta) + (1-\lambda)I_{y|\mathbf{x}}(\theta) ,$$

$$V_\lambda(\theta) = \text{Cov}\left(\frac{\partial \ell_\lambda(\theta)}{\partial\theta}\right) = E\left\{\left(\frac{\partial \ell_\lambda(\theta)}{\partial\theta}\right)^2\right\} = \lambda^2 I_{\text{g}}(\theta) + (1-\lambda^2)I_{y|\mathbf{x}}(\theta) .$$

Here, after some algebra, we obtain

$$\frac{1}{n}I_{y|\mathbf{x}}(\theta) = \int_{\mathbf{x}} p(\mathcal{C}_1|\mathbf{x})p(\mathcal{C}_0|\mathbf{x})\left[\frac{\partial \log r(\theta,\pi;\mathbf{x})}{\partial\theta}\right]\left[\frac{\partial \log r(\theta,\pi;\mathbf{x})}{\partial\theta}\right]^T p(\mathbf{x})d\mathbf{x} ,$$

with $r(\theta,\pi;\mathbf{x}) = \frac{\pi_1 p(\mathbf{x}|\theta_1)}{\pi_0 p(\mathbf{x}|\theta_0)}$ and $p(\mathbf{x}) = \pi_1 p(\mathbf{x}|\theta_1) + \pi_0 p(\mathbf{x}|\theta_0)$.

Lemma A.1.1 *When $\lambda = 1$, we have $I_\lambda(\theta) = V_\lambda(\theta) = I_g(\theta)$, and thus $\Sigma_\lambda(\hat{\theta}) = nI_g^{-1}(\theta)$; when $\lambda = 0$, we have $I_\lambda(\theta) = V_\lambda(\theta) = I_{y|\mathbf{x}}(\theta)$, and thus $\Sigma_\lambda(\hat{\theta}) = nI_{y|\mathbf{x}}^{-1}(\theta)$.* ∎

With regard to each component of θ, we obtain

$$\frac{\partial \log r(\theta,\pi;\mathbf{x})}{\partial\pi_1} = \frac{1}{\pi_1\pi_0} ,$$

$$\frac{\partial \log r(\theta,\pi;\mathbf{x})}{\partial\mu_1} = W(\mathbf{x}-\mu_1) , \ \frac{\partial \log r(\theta,\pi;\mathbf{x})}{\partial\mu_0} = -W(\mathbf{x}-\mu_0) ,$$

$$\left[\frac{\partial \log r(\theta,\pi;\mathbf{x})}{\partial W}\right]_{i,j} = \frac{\left[-(\mathbf{x}-\mu_1)(\mathbf{x}-\mu_1)^T + (\mathbf{x}-\mu_0)(\mathbf{x}-\mu_0)^T\right]_{i,j}}{1+\delta_{i,j}} .$$

A.1.4 Relationship between $d\alpha = (\hat{\alpha} - \alpha)$ and $d\theta = (\hat{\theta} - \theta)$

With $g(\mathbf{x}, \alpha) = \log \frac{\pi_1}{\pi_0} - \frac{1}{2}(\mu_1^T W \mu_1 - \mu_0^T W \mu_0) + (\mu_1 - \mu_0)^T W \mathbf{x} = \beta_0 + \beta^T \mathbf{x}$, after some algebra, we obtain

$$\frac{\partial \beta_0}{\partial \pi_1} = \frac{1}{\pi_1 \pi_0} \; , \; \frac{\partial \beta_0}{\partial \mu_1^T} = -\mu_1^T W \; , \; \frac{\partial \beta_0}{\partial \mu_0^T} = \mu_0^T W \; , \; \frac{\partial \beta_0}{\partial W_{i,j}} = \frac{\left[-\mu_1 \mu_1^T + \mu_0 \mu_0^T\right]_{i,j}}{1 + \delta_{i,j}} \; ,$$

$$\frac{\partial \beta}{\partial \pi_1} = \mathbf{0} \; , \; \frac{\partial \beta}{\partial \mu_1^T} = W \; , \; \frac{\partial \beta}{\partial \mu_0^T} = -W \; , \; \frac{\partial \beta}{\partial W_{i,j}} = \frac{\left[\mathbf{J}_{i,j} + \mathbf{J}_{j,i}\right](\mu_1 - \mu_0)}{1 + \delta_{i,j}} \; ,$$

where $\mathbf{J}_{i,j}$ is the single-entry matrix with 0 everywhere except for 1 at the (i, j)-th position.

Using the above differentiation results, combined with $\Sigma_{\mathrm{g}}(\hat{\theta})$ as derived in Section A.1.2 and $\Sigma_{\lambda}(\hat{\theta})$ as derived in Section A.1.3, we can obtain the $(p+1) \times (p+1)$ matrices $\Sigma_{\mathrm{g}}(\hat{\alpha})$ and $\Sigma_{\lambda}(\hat{\alpha})$, respectively.

A.1.5 Estimation of $\Sigma_{\mathrm{d}}(\hat{\alpha})$

As mentioned earlier in Section A.1.3, for the discriminative component in the GDT, we have

$$\frac{1}{n} I_{y|\mathbf{x}}(\theta) = \int_{\mathbf{x}} p(\mathcal{C}_1|\mathbf{x}) p(\mathcal{C}_0|\mathbf{x}) \left[\frac{\partial \log r(\theta, \pi; \mathbf{x})}{\partial \theta}\right] \left[\frac{\partial \log r(\theta, \pi; \mathbf{x})}{\partial \theta}\right]^T p(\mathbf{x}) d\mathbf{x} \; .$$

Similarly, for discriminative learning of the LLR estimator $\hat{\alpha}$, its asymptotic variance matrix $\Sigma_{\mathrm{d}}(\hat{\alpha})$ was proved by O'Neill (1980) to be

$$\begin{aligned} \Sigma_{\mathrm{d}}^{-1}(\hat{\alpha}) &= \int_{\mathbf{x}} p(\mathcal{C}_1|\mathbf{x}) p(\mathcal{C}_0|\mathbf{x}) \left[\frac{\partial g(\mathbf{x}, \alpha)}{\partial \alpha}\right] \left[\frac{\partial g(\mathbf{x}, \alpha)}{\partial \alpha}\right]^T p(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathbf{x}} \frac{e^{g(\mathbf{x}, \alpha)}}{[1 + e^{g(\mathbf{x}, \alpha)}]^2} \begin{pmatrix} 1 \\ \mathbf{x} \end{pmatrix} \left(1 \; \mathbf{x}^T\right) p(\mathbf{x}) d\mathbf{x} \; . \end{aligned}$$

A.1.6 Estimation of $\mathbf{B}$

To calculate AER and ARE, such as

$$\mathrm{ARE}(\hat{\alpha}_{\mathrm{d}}, \hat{\alpha}_{\mathrm{g}}) = \frac{\mathrm{tr}(\mathbf{B} \Sigma_{\mathrm{g}}(\hat{\alpha}))}{\mathrm{tr}(\mathbf{B} \Sigma_{\mathrm{d}}(\hat{\alpha}))} \; ,$$

we need to derive $\mathbf{B}$, which was defined in Section 3.2.1.

For $g(\mathbf{x}, \alpha) = \beta_0 + \beta^T \mathbf{x}$, $\mathbf{x}|\theta_1 \sim \mathcal{N}(\mu_1, \Sigma)$, $\mathbf{x}|\theta_0 \sim \mathcal{N}(\mu_0, \Sigma)$, we have

$$\mathbf{B} = \frac{1}{4\sqrt{\beta^T \beta}} \int_D \begin{pmatrix} 1 \\ \mathbf{x} \end{pmatrix} \left(1 \; \mathbf{x}^T\right) p(\mathbf{x}) dm_D \; ,$$

where $D = \{\mathbf{x} : g(\mathbf{x}, \alpha) = 0\}$ and m_D is Lebesgue measure on D.

A.1.7 Simplified Estimation by Linear Transformation of $\mathbf{x}$

Since a linear transformation of $\mathbf{x}$ into $a + A\mathbf{x}$ does not change the misclassification error rates, the above mentioned estimation of asymptotic variance matrices can be simplified by a workable transformation. (Hereafter we still use $\mathbf{x}$ to denote the new feature vector obtained from transformation.)

Efron (1975) suggested a new, linearly transformed $\mathbf{x}$ satisfying: $\mathbf{x}|\theta_1 \sim \mathcal{N}(\frac{\Delta}{2}\mathbf{e}_1, \mathbf{I})$, $\mathbf{x}|\theta_0 \sim \mathcal{N}(-\frac{\Delta}{2}\mathbf{e}_1, \mathbf{I})$, where $\Delta = \sqrt{(\mu_1 - \mu_0)^T W (\mu_1 - \mu_0)}$, the Mahalanobis distance between the means of the two sub-populations, and, in addition, it is required that $\Delta \neq 0$ to make the two sub-populations nonidentical; $\mathbf{I}$ is the identity matrix and $\mathbf{e}_1^T = (1, 0, 0, \ldots, 0)$. In such a case,

$$\frac{\partial \beta_0}{\partial W_{i,j}} = \frac{\left[-\mu_1\mu_1^T + \mu_0\mu_0^T\right]_{i,j}}{1 + \delta_{i,j}} = 0, \ \frac{\partial \beta}{\partial W_{i,j}} = \frac{\Delta\left[\mathbf{J}_{i,j} + \mathbf{J}_{j,i}\right]\mathbf{e}_1}{1 + \delta_{i,j}} \ .$$

This suggests separating $(\text{vech}(W))^T$ into (η_1^T, η_2^T), where $\eta_1^T = (W_{1,1}, W_{1,2}, \ldots, W_{1,p})$ and $\eta_2^T = (W_{2,2}, W_{2,3}, \ldots, W_{p,p})$, so that, after some algebra, $\frac{\partial \beta}{\partial \eta_1} = \Delta\mathbf{I}$, $\frac{\partial \beta}{\partial \eta_2} = \mathbf{0}$.

Through simplification, we obtain $d\alpha = M d\theta$, where $M = \begin{pmatrix} \frac{1}{\pi_1\pi_0} & -\frac{\Delta}{2}\mathbf{e}_1^T & -\frac{\Delta}{2}\mathbf{e}_1^T & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & -\mathbf{I} & \Delta\mathbf{I} & \mathbf{0} \end{pmatrix}$.
Since the last column of the block matrix M is all zeros, and all the components of θ are asymptotically uncorrelated, we can ignore the asymptotic covariance matrix of the vector η_2 for the computation of $\Sigma_{\text{g}}(\hat{\alpha})$ and $\Sigma_{\lambda}(\hat{\alpha})$.

A.1.7.1 Re-calculation of $I_{\text{g}}(\theta)$, $\Sigma_{\text{g}}(\hat{\theta})$ and $\Sigma_{\text{g}}(\hat{\alpha})$

If $\mathbf{x}|\theta_1 \sim \mathcal{N}(\frac{\Delta}{2}\mathbf{e}_1, \mathbf{I})$ and $\mathbf{x}|\theta_0 \sim \mathcal{N}(-\frac{\Delta}{2}\mathbf{e}_1, \mathbf{I})$, we can obtain
$\sqrt{n}(\hat{\pi}_1 - \pi_1) \sim \mathcal{AN}(0, \pi_1\pi_0)$,
$\sqrt{n}(\hat{\mu}_1 - \mu_1) \sim \mathcal{AN}(\mathbf{0}, \frac{1}{\pi_1}\mathbf{I})$, $\sqrt{n}(\hat{\mu}_0 - \mu_0) \sim \mathcal{AN}(\mathbf{0}, \frac{1}{\pi_0}\mathbf{I})$,
$\sqrt{n}(\hat{\eta}_1 - \eta_1) \sim \mathcal{AN}(\mathbf{0}, nI_{\text{g}}^{-1}(\eta_1))$, where $\left[\frac{1}{n}I_{\text{g}}(\eta_1)\right]_{j,l} = \frac{\mathbf{I}_{1,1}\mathbf{I}_{l,j} + \mathbf{I}_{1,l}\mathbf{I}_{1,j}}{(1+\delta_{1,j})(1+\delta_{1,l})}$, so that $\sqrt{n}(\hat{\eta}_1 - \eta_1) \sim \mathcal{AN}(\mathbf{0}, \mathbf{J}_{1,1} + \mathbf{I})$.
It then follows that

$$\Sigma_{\text{g}}(\hat{\theta}) = \text{Block-Diag}(\Sigma_{\text{g}}(\hat{\pi}_1), \Sigma_{\text{g}}(\hat{\mu}_1), \Sigma_{\text{g}}(\hat{\mu}_0), \Sigma_{\text{g}}(\hat{\eta}_1), \Sigma_{\text{g}}(\hat{\eta}_2))$$

$$= \text{Block-Diag}(\pi_1\pi_0, \frac{1}{\pi_1}\mathbf{I}, \frac{1}{\pi_0}\mathbf{I}, \mathbf{J}_{1,1} + \mathbf{I}, \pounds) \ ,$$

where $\pounds$ is ignored, and the $(p+1)\times(p+1)$ symmetric matrix $\Sigma_{\rm g}(\hat{\alpha})$ is (Efron, 1975)

$$\Sigma_{\rm g}(\hat{\alpha}) = M\Sigma_{\rm g}(\hat{\theta})M^T = \begin{bmatrix} [\Sigma_{\rm g}(\hat{\alpha})]_{1,1} & [\Sigma_{\rm g}(\hat{\alpha})]_{1,2} & \\ [\Sigma_{\rm g}(\hat{\alpha})]_{2,1} & [\Sigma_{\rm g}(\hat{\alpha})]_{2,2} & \\ & & [\Sigma_{\rm g}(\hat{\alpha})]_{3,3}\mathbf{I}_{p-1} \end{bmatrix}$$

$$= \frac{1}{\pi_1\pi_0}\begin{bmatrix} 1+\frac{\Delta^2}{4} & \frac{\Delta(\pi_1-\pi_0)}{2} & \\ \frac{\Delta(\pi_1-\pi_0)}{2} & 1+2\Delta^2\pi_1\pi_0 & \\ & & (1+\Delta^2\pi_1\pi_0)\mathbf{I}_{p-1} \end{bmatrix}.$$

Lemma A.1.2 *When $\pi_1=\pi_0=\frac{1}{2}$, we have a diagonal matrix $\Sigma_g(\hat{\alpha}) = Diag(4+\Delta^2, 4+2\Delta^2, (4+\Delta^2)\mathbf{I}_{p-1})$;* i.e., *in this case, there exists a linear transformation of $\mathbf{x}$ that can make the generative estimates $\hat{\alpha}=\hat{\alpha}_g$ of the coefficients of the linear discriminant function $g(\mathbf{x},\alpha)$ asymptotically uncorrelated.* ∎

A.1.7.2 Re-calculation of $I_{y|\mathbf{x}}(\theta)$, $\Sigma_\lambda(\hat{\theta})$ and $\Sigma_\lambda(\hat{\alpha})$

After some algebra, we can obtain

$\frac{1}{n}I_{y|\mathbf{x}}(\pi_1) = \frac{A_0}{\pi_1\pi_0}$,

$\frac{1}{n}I_{y|\mathbf{x}}(\mu_1) = \pi_1\pi_0\text{Diag}(A_2 - \Delta A_1 + \frac{\Delta^2}{4}A_0, A_0, \ldots, A_0)$,

$\frac{1}{n}I_{y|\mathbf{x}}(\mu_0) = \pi_1\pi_0\text{Diag}(A_2 + \Delta A_1 + \frac{\Delta^2}{4}A_0, A_0, \ldots, A_0)$,

$\frac{1}{n}I_{y|\mathbf{x}}(\eta_1) = \pi_1\pi_0\Delta^2\text{Diag}(A_2, A_0, \ldots, A_0)$,

where, with $\phi(x)$ denoting the density of the univariate standard normal distribution,

$$A_i = \int_{-\infty}^{\infty} \frac{e^{-\frac{\Delta^2}{8}}x^i\phi(x)}{\pi_1 e^{\frac{\Delta}{2}x} + \pi_0 e^{-\frac{\Delta}{2}x}}dx,\ i=0,1,\ldots.$$

Lemma A.1.3 *For all $k=0,1,\ldots$, $A_{2k}\geq 0$, and A_{2k} is even-symmetric while A_{2k+1} is odd-symmetric about $\pi_1=\frac{1}{2}$ (so that $A_{2k+1}=0$ if $\pi_1=\pi_0=\frac{1}{2}$).*

Lemma A.1.4 *When $\Delta\to 0$, we have that A_i, $i=0,1,\ldots$, is the i-th moment of the univariate standard normal distribution $\mathcal{N}(0,1)$ so that $A_0=1$, $A_1=0$, $A_2=1,\ldots$.* ∎

With $I_{y|\mathbf{x}}(\theta)$ and $I_{\rm g}(\theta)$ as given in Section A.1.7.1, we can first derive $I_\lambda(\theta)$ and $V_\lambda(\theta)$ through

$$\begin{cases} I_\lambda(\theta) = \lambda I_{\rm g}(\theta) + (1-\lambda)I_{y|\mathbf{x}}(\theta) \\ V_\lambda(\theta) = \lambda^2 I_{\rm g}(\theta) + (1-\lambda^2)I_{y|\mathbf{x}}(\theta)\ , \end{cases}$$

and then derive $\Sigma_\lambda(\hat{\theta})$ and $\Sigma_\lambda(\hat{\alpha})$ through

$$\begin{cases} \Sigma_\lambda(\hat{\theta}) = nI_\lambda^{-1}(\theta)V_\lambda(\theta)I_\lambda^{-1}(\theta) \\ \Sigma_\lambda(\hat{\alpha}) = M\Sigma_\lambda(\hat{\theta})M^T \ . \end{cases}$$

More precisely, $\Sigma_\lambda(\hat{\theta})$ is a block-diagonal matrix composed of a scalar $\Sigma_\lambda(\hat{\pi}_1)$, three $p \times p$ diagonal matrices $\Sigma_\lambda(\hat{\mu}_1)$, $\Sigma_\lambda(\hat{\mu}_0)$ and $\Sigma_\lambda(\hat{\eta}_1)$, and a matrix of no interest $\Sigma_\lambda(\hat{\eta}_2)$, where

$$\Sigma_\lambda(\hat{\pi}_1) = \pi_1\pi_0\frac{\lambda^2+(1-\lambda^2)A_0}{[\lambda+(1-\lambda)A_0]^2} \ ,$$

$$\Sigma_\lambda(\hat{\mu}_1) = \begin{bmatrix} [\Sigma_\lambda(\hat{\mu}_1)]_{1,1} & \\ & [\Sigma_\lambda(\hat{\mu}_1)]_{2,2}\mathbf{I}_{p-1} \end{bmatrix} = \frac{1}{\pi_1}\begin{bmatrix} \frac{\lambda^2+(1-\lambda^2)\pi_0(A_2-\Delta A_1+\frac{\Delta^2}{4}A_0)}{[\lambda+(1-\lambda)\pi_0(A_2-\Delta A_1+\frac{\Delta^2}{4}A_0)]^2} & \\ & \frac{\lambda^2+(1-\lambda^2)\pi_0 A_0}{[\lambda+(1-\lambda)\pi_0 A_0]^2}\mathbf{I}_{p-1} \end{bmatrix} ,$$

$$\Sigma_\lambda(\hat{\mu}_0) = \begin{bmatrix} [\Sigma_\lambda(\hat{\mu}_0)]_{1,1} & \\ & [\Sigma_\lambda(\hat{\mu}_0)]_{2,2}\mathbf{I}_{p-1} \end{bmatrix} = \frac{1}{\pi_0}\begin{bmatrix} \frac{\lambda^2+(1-\lambda^2)\pi_1(A_2+\Delta A_1+\frac{\Delta^2}{4}A_0)}{[\lambda+(1-\lambda)\pi_1(A_2+\Delta A_1+\frac{\Delta^2}{4}A_0)]^2} & \\ & \frac{\lambda^2+(1-\lambda^2)\pi_1 A_0}{[\lambda+(1-\lambda)\pi_1 A_0]^2}\mathbf{I}_{p-1} \end{bmatrix} ,$$

and

$$\Sigma_\lambda(\hat{\eta}_1) = \begin{bmatrix} [\Sigma_\lambda(\hat{\eta}_1)]_{1,1} & \\ & [\Sigma_\lambda(\hat{\eta}_1)]_{2,2}\mathbf{I}_{p-1} \end{bmatrix} = \begin{bmatrix} \frac{\frac{1}{2}\lambda^2+(1-\lambda^2)\pi_0\pi_1\Delta^2 A_2}{[\frac{1}{2}\lambda+(1-\lambda)\pi_0\pi_1\Delta^2 A_2]^2} & \\ & \frac{\lambda^2+(1-\lambda^2)\pi_0\pi_1\Delta^2 A_0}{[\lambda+(1-\lambda)\pi_0\pi_1\Delta^2 A_0]^2}\mathbf{I}_{p-1} \end{bmatrix} .$$

Therefore, we have

$$\Sigma_\lambda(\hat{\alpha}) = \begin{bmatrix} [\Sigma_\lambda(\hat{\alpha})]_{1,1} & [\Sigma_\lambda(\hat{\alpha})]_{1,2} & \\ [\Sigma_\lambda(\hat{\alpha})]_{2,1} & [\Sigma_\lambda(\hat{\alpha})]_{2,2} & \\ & & [\Sigma_\lambda(\hat{\alpha})]_{3,3}\mathbf{I}_{p-1} \end{bmatrix} ,$$

where

$$[\Sigma_\lambda(\hat{\alpha})]_{1,1} = (\frac{1}{\pi_0\pi_1})^2\Sigma_\lambda(\hat{\pi}_1) + \frac{\Delta^2}{4}([\Sigma_\lambda(\hat{\mu}_1)]_{1,1} + [\Sigma_\lambda(\hat{\mu}_0)]_{1,1}) \ ,$$

$$[\Sigma_\lambda(\hat{\alpha})]_{1,2} = [\Sigma_\lambda(\hat{\alpha})]_{2,1} = \frac{\Delta}{2}(-[\Sigma_\lambda(\hat{\mu}_1)]_{1,1} + [\Sigma_\lambda(\hat{\mu}_0)]_{1,1}) \ ,$$

$$[\Sigma_\lambda(\hat{\alpha})]_{2,2} = [\Sigma_\lambda(\hat{\mu}_1)]_{1,1} + [\Sigma_\lambda(\hat{\mu}_0)]_{1,1} + \Delta^2[\Sigma_\lambda(\hat{\eta}_1)]_{1,1} \ ,$$

$$[\Sigma_\lambda(\hat{\alpha})]_{3,3} = [\Sigma_\lambda(\hat{\mu}_1)]_{2,2} + [\Sigma_\lambda(\hat{\mu}_0)]_{2,2} + \Delta^2[\Sigma_\lambda(\hat{\eta}_1)]_{2,2} \ .$$

Lemma A.1.5 *When $\pi_1 = \pi_0 = \frac{1}{2}$, according to Lemma A.1.3, we have $\Sigma_\lambda(\hat{\mu}_1) = \Sigma_\lambda(\hat{\mu}_0)$ and thus $[\Sigma_\lambda(\hat{\alpha})]_{1,2} = [\Sigma_\lambda(\hat{\alpha})]_{2,1} = 0$, leading to a diagonal matrix $\Sigma_\lambda(\hat{\alpha})$;* i.e., *in this case, there exists a linear transformation of $\mathbf{x}$ that can make the GDT estimates $\hat{\alpha} = \hat{\alpha}_\lambda$ asymptotically uncorrelated.* ∎

A.1.7.3 Re-calculation of $\Sigma_{\mathrm{d}}(\hat{\alpha})$

Efron (1975) showed that

$$\Sigma_{\mathrm{d}}^{-1}(\hat{\alpha}) = \pi_1\pi_0 \begin{bmatrix} A_0 & A_1 & \\ A_1 & A_2 & \\ & & A_0\mathbf{I}_{p-1} \end{bmatrix},$$

where A_i is as defined in Section A.1.7.2. It follows that

$$\Sigma_{\mathrm{d}}(\hat{\alpha}) = \begin{bmatrix} [\Sigma_{\mathrm{d}}(\hat{\alpha})]_{1,1} & [\Sigma_{\mathrm{d}}(\hat{\alpha})]_{1,2} & \\ [\Sigma_{\mathrm{d}}(\hat{\alpha})]_{2,1} & [\Sigma_{\mathrm{d}}(\hat{\alpha})]_{2,2} & \\ & & [\Sigma_{\mathrm{d}}(\hat{\alpha})]_{3,3}\mathbf{I}_{p-1} \end{bmatrix} = \frac{1}{\pi_1\pi_0} \begin{bmatrix} \frac{A_2}{A_0A_2-A_1^2} & \frac{-A_1}{A_0A_2-A_1^2} & \\ \frac{-A_1}{A_0A_2-A_1^2} & \frac{A_0}{A_0A_2-A_1^2} & \\ & & \frac{1}{A_0}\mathbf{I}_{p-1} \end{bmatrix}.$$

Lemma A.1.6 *When $\pi_1 = \pi_0 = \frac{1}{2}$, according to Lemma A.1.3, we have a diagonal matrix $\Sigma_g(\hat{\alpha}) = Diag(\frac{4}{A_0}, \frac{4}{A_2}, \frac{4}{A_0}\mathbf{I}_{p-1})$;* i.e., *in this case, there exists a linear transformation of $\mathbf{x}$ that can make the discriminative estimates $\hat{\alpha} = \hat{\alpha}_d$ asymptotically uncorrelated.* ∎

A.1.7.4 Re-calculation of B

For $g(\mathbf{x}, \alpha) = \beta_0 + \beta^T\mathbf{x}$, $\mathbf{x}|\theta_1 \sim \mathcal{N}(\frac{\Delta}{2}\mathbf{e}_1, \mathbf{I})$, $\mathbf{x}|\theta_0 \sim \mathcal{N}(-\frac{\Delta}{2}\mathbf{e}_1, \mathbf{I})$, $\Delta > 0$, we have that, after some algebra, $\beta_0 = \log\frac{\pi_1}{\pi_0}$, $\beta^T = \Delta\mathbf{e}_1^T$, so $\beta^T\beta = \Delta^2$, and
$g(\tilde{\mathbf{x}}, \alpha) = 0 \Leftrightarrow \tilde{\mathbf{x}}^T = (\tau = -\frac{1}{\Delta}\log\frac{\pi_1}{\pi_0}, x_2, \ldots, x_p)$, where $x_2, \ldots, x_p$ are any real numbers. It follows that

$$\mathbf{B} = \frac{\pi_1\phi(\tau - \frac{\Delta}{2})}{2\Delta} \begin{bmatrix} 1 & \tau & \\ \tau & \tau^2 & \\ & & \mathbf{I}_{p-1} \end{bmatrix}.$$

A.2 Asymptotic Efficiency of GDT for Quadratic Normal Discrimination

A.2.1 Quadratic Normal Discrimination

Now we assume that, within each sub-population, the feature vector $\mathbf{x}$ arises from one of two multivariate normal distributions with different covariance matrices, *i.e.*, $\mathbf{x}|\theta_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$, $\mathbf{x}|\theta_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$, where $\Sigma_1 \neq \Sigma_0$. In addition, no mis-specification occurs. In this context, a quadratic discriminant function is derived, as in Section 1.1.4, to be

$$g(\mathbf{x}, \alpha) = \beta_0 + \beta^T\mathbf{x} + \mathbf{x}^T\Gamma\mathbf{x} =$$

$$\log\frac{\pi_1}{\pi_0}-\frac{1}{2}(\mu_1^T W_1\mu_1-\mu_0^T W_0\mu_0)-\frac{1}{2}\log\frac{|W_0|}{|W_1|}+(\mu_1^T W_1-\mu_0^T W_0)\mathbf{x}-\frac{1}{2}\mathbf{x}^T(W_1-W_0)\mathbf{x}\,,$$

where $W_1=\Sigma_1^{-1}$, $W_0=\Sigma_0^{-1}$, so that Γ is a symmetric matrix, $\alpha^T=(\beta_0,\beta^T,(\text{vech}(\Gamma))^T)$ and $\theta^T=(\pi_1,\mu_1^T,\mu_0^T,(\text{vech}(W_1))^T,(\text{vech}(W_0))^T)$.

A.2.2 Estimation of $\Sigma_{\text{g}}(\hat{\theta})$

By calculating the Fisher information matrix and after some algebra, as with linear normal discrimination, we obtain that

$\sqrt{n}(\hat{\pi}_1-\pi_1)\sim\mathcal{AN}(0,\pi_1\pi_0)$,

$\sqrt{n}(\hat{\mu}_y-\mu_y)\sim\mathcal{AN}(\mathbf{0},\frac{1}{\pi_y}\Sigma_y), y=0,1$,

$\sqrt{n}(\text{vech}(\hat{W}_y)-\text{vech}(W_y))\sim\mathcal{AN}(\mathbf{0},nI_{\text{g}}^{-1}(\text{vech}(W_y)))$, where $y=0,1$, where

$$\left\{I_{\text{g}}(\text{vech}(W_y))\right\}_{[W_y]_{i,j},[W_y]_{k,l}}=E\left\{-\frac{\partial^2\ell_{\text{g}}(\theta)}{\partial[W_y]_{i,j}\partial[W_y]_{k,l}}\right\}$$

$$=\pi_y\frac{n([\Sigma_y]_{i,k}[\Sigma_y]_{l,j}+[\Sigma_y]_{i,l}[\Sigma_y]_{k,j})}{(1+\delta_{i,j})(1+\delta_{k,l})}\,.$$

It follows that $\Sigma_{\text{g}}(\hat{\theta})$ is a block-diagonal matrix composed of a scalar $\Sigma_{\text{g}}(\hat{\pi}_1)=\pi_1\pi_0$, two $p\times p$ matrices $\Sigma_{\text{g}}(\hat{\mu}_1)=\frac{1}{\pi_1}\Sigma_1$ and $\Sigma_{\text{g}}(\hat{\mu}_0)=\frac{1}{\pi_0}\Sigma_0$, and two $\frac{p(p+1)}{2}\times\frac{p(p+1)}{2}$ matrices $\Sigma_{\text{g}}(\text{vech}(\hat{W}_1))=nI_{\text{g}}^{-1}(\text{vech}(W_1))$ and $\Sigma_{\text{g}}(\text{vech}(\hat{W}_0))=nI_{\text{g}}^{-1}(\text{vech}(W_0))$.

A.2.3 Estimation of $\Sigma_\lambda(\hat{\theta})$

The way to estimate $\Sigma_\lambda(\hat{\theta})$ is similar to that in Section A.1.3, based on the calculation of $I_\lambda(\theta)$ and $V_\lambda(\theta)$, or, more concretely, on the calculation of $I_{\text{g}}(\theta)$ (see Section A.2.2) and $I_{y|\mathbf{x}}(\theta)$. In order to calculate $I_{y|\mathbf{x}}(\theta)$, we derive

$$\frac{\partial\log r(\theta,\pi;\mathbf{x})}{\partial\pi_1}=\frac{1}{\pi_1\pi_0}\,,$$

$$\frac{\partial\log r(\theta,\pi;\mathbf{x})}{\partial\mu_1}=W_1(\mathbf{x}-\mu_1),\ \frac{\partial\log r(\theta,\pi;\mathbf{x})}{\partial\mu_0}=-W_0(\mathbf{x}-\mu_0)\,,$$

$$\left[\frac{\partial\log r(\theta,\pi;\mathbf{x})}{\partial W_y}\right]_{i,j}=\frac{\left[-(\mathbf{x}-\mu_y)(\mathbf{x}-\mu_y)^T+\Sigma_y\right]_{i,j}}{1+\delta_{i,j}}(-1)^{1-y}, y=0,1\,.$$

A.2.4 Relationship between $d\alpha = (\hat{\alpha} - \alpha)$ and $d\theta = (\hat{\theta} - \theta)$

Considering $\beta_0 = \log \frac{\pi_1}{\pi_0} - \frac{1}{2}(\mu_1^T W_1 \mu_1 - \mu_0^T W_0 \mu_0) - \frac{1}{2} \log \frac{|W_0|}{|W_1|}$, $\beta^T = \mu_1^T W_1 - \mu_0^T W_0$ and $\Gamma = -\frac{1}{2}(W_1 - W_0)$, after some algebra, we obtain that

$$\frac{\partial \beta_0}{\partial \pi_1} = \frac{1}{\pi_1 \pi_0} \, , \; \frac{\partial \beta_0}{\partial \mu_1^T} = -\mu_1^T W_1 \, , \; \frac{\partial \beta_0}{\partial \mu_0^T} = \mu_0^T W_0 \, ,$$

$$\frac{\partial \beta_0}{\partial [W_1]_{i,j}} = \frac{\left[-\mu_1 \mu_1^T + \Sigma_1\right]_{i,j}}{1 + \delta_{i,j}} \, , \; \frac{\partial \beta_0}{\partial [W_0]_{i,j}} = -\frac{\left[-\mu_0 \mu_0^T + \Sigma_0\right]_{i,j}}{1 + \delta_{i,j}} \, ,$$

$$\frac{\partial \beta}{\partial \pi_1} = \mathbf{0} \, , \; \frac{\partial \beta}{\partial \mu_1^T} = W_1 \, , \; \frac{\partial \beta}{\partial \mu_0^T} = -W_0 \, ,$$

$$\frac{\partial \beta}{\partial [W_1]_{i,j}} = \frac{[\mathbf{J}_{i,j} + \mathbf{J}_{j,i}]\,(\mu_1)}{1 + \delta_{i,j}} \, , \; \frac{\partial \beta}{\partial [W_0]_{i,j}} = \frac{[\mathbf{J}_{i,j} + \mathbf{J}_{j,i}]\,(-\mu_0)}{1 + \delta_{i,j}} \, ,$$

and

$$\frac{\partial \Gamma_{i,j}}{\partial \pi_1} = \mathbf{0} \, , \; \frac{\partial \Gamma_{i,j}}{\partial \mu_1^T} = \mathbf{0} \, , \; \frac{\partial \Gamma_{i,j}}{\partial \mu_0^T} = \mathbf{0} \, ,$$

$$\frac{\partial \Gamma}{\partial [W_1]_{i,j}} = -\frac{1}{2} \frac{[\mathbf{J}_{i,j} + \mathbf{J}_{j,i}]}{1 + \delta_{i,j}} \, , \; \frac{\partial \Gamma}{\partial [W_0]_{i,j}} = \frac{1}{2} \frac{[\mathbf{J}_{i,j} + \mathbf{J}_{j,i}]}{1 + \delta_{i,j}} \, .$$

Using the above differentiation results, combined with $\Sigma_{\text{g}}(\hat{\theta})$ as derived in Section A.2.2 and $\Sigma_{\lambda}(\hat{\theta})$ as derived in Section A.2.3, we can obtain $\Sigma_{\text{g}}(\hat{\alpha})$ and $\Sigma_{\lambda}(\hat{\alpha})$.

A.2.5 Estimation of $\Sigma_{\text{d}}(\hat{\alpha})$

Similarly to that in Section A.1.5, the asymptotic variance matrix $\Sigma_{\text{d}}(\hat{\alpha})$ for quadratic normal discrimination is

$$\Sigma_{\text{d}}^{-1}(\hat{\alpha}) = \int_{\mathbf{x}} \frac{e^{g(\mathbf{x},\alpha)}}{[1 + e^{g(\mathbf{x},\alpha)}]^2} [\nabla_{\alpha} g(\mathbf{x}, \alpha)][\nabla_{\alpha} g(\mathbf{x}, \alpha)]^T p(\mathbf{x}) d\mathbf{x} \, ,$$

where $\nabla_{\alpha} g(\mathbf{x}, \alpha) = \left(1 \; \mathbf{x}^T \; \mathbf{s_x}^T\right)^T$, in which $\mathbf{s_x} = \text{vech}(2\mathbf{x}\mathbf{x}^T - \text{Diag}(\mathbf{x}\mathbf{x}^T))$.

A.2.6 Estimation of $\mathbf{B}$

For $g(\mathbf{x}, \alpha) = \beta_0 + \beta^T \mathbf{x} + \mathbf{x}^T \Gamma \mathbf{x}$, $\mathbf{x}|\theta_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$, $\mathbf{x}|\theta_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$, we have

$$|\nabla_{\mathbf{x}} g(\mathbf{x}, \alpha)|^2 = \sum_{k=1}^{p} \left(\frac{\partial g}{\partial \mathbf{x}_k} \right)^2 = \sum_{k=1}^{p} \left(\beta_k + 2\mathbf{x}_k \gamma_{k,k} + \sum_{i=1, i \neq k}^{p} 2\mathbf{x}_i \gamma_{k,i} \right)^2 \, .$$

We may then calculate $\mathbf{B}$ based on its definition in Section 3.2.1.

A.2.7 Simplified Estimation by Linear Transformation of $\mathbf{x}$

Here we consider a univariate case used by O'Neill (1980), *i.e.*, assuming $p = 1$, $x|\theta_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $x|\theta_0 \sim \mathcal{N}(\mu_0, \sigma_0^2)$, so that
$\theta^T = (\pi_1, \mu_1, \mu_0, \eta_1, \eta_0)$, where $\eta_1 = 1/\sigma_1^2, \eta_0 = 1/\sigma_0^2$ and $\pi_1 \in (0, 1)$
and $g(x, \alpha) = \beta_0 + \beta x + \gamma x^2$, $\alpha^T = (\beta_0, \beta, \gamma)$.

Furthermore, O'Neill (1980) suggested a linearly transformed x satisfying $x|\theta_1 \sim \mathcal{N}(\mu, 1)$, $x|\theta_0 \sim \mathcal{N}(0, \rho)$, $\rho < 1$, which may further simplify the computation. However, the following derivations in this paper are valid for $0 < \rho < 1$ and $\rho > 1$; note that $\rho \neq 1$ is necessary to prevent the quadratic normal discrimination from degenerating into linear normal discrimination, which has been discussed in Section A.1.

Through the simplification, for $d\alpha = M d\theta$, we have

$$M = \begin{pmatrix} \frac{1}{\pi_1\pi_0} & -\mu_1\eta_1 & \mu_0\eta_0 & \frac{-\mu_1^2+\frac{1}{\eta_1}}{2} & \frac{\mu_0^2-\frac{1}{\eta_0}}{2} \\ 0 & \eta_1 & -\eta_0 & \mu_1 & -\mu_0 \\ 0 & 0 & 0 & -\frac{1}{2} & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\pi_1\pi_0} & -\mu & 0 & \frac{-\mu^2+1}{2} & -\frac{\rho}{2} \\ 0 & 1 & -\frac{1}{\rho} & \mu & 0 \\ 0 & 0 & 0 & -\frac{1}{2} & \frac{1}{2} \end{pmatrix} .$$

In addition, since the distributions of $x|\theta_1$ and $x|\theta_0$ are symmetric about their corresponding means, μ_1 and μ_0, respectively, it is expected that an index of misclassification error, such as AER and ARE, ought to be invariant either to the symmetric change of μ_1 about μ_0 into $\mu_1' = 2\mu_0 - \mu_1$ or to the symmetric change of μ_0 about μ_1 into $\mu_0' = 2\mu_1 - \mu_0$. After the above-mentioned simplification, as a specific instance, it can be illustrated that both AER and ARE are invariant to the symmetric change of μ into $\mu' = -\mu$.

A.2.7.1 Re-calculation of $I_{\mathrm{g}}(\theta)$, $\Sigma_{\mathrm{g}}(\hat{\theta})$ and $\Sigma_{\mathrm{g}}(\hat{\alpha})$

Considering $x|\theta_1 \sim \mathcal{N}(\mu, 1)$, $x|\theta_0 \sim \mathcal{N}(0, \rho)$, $\theta^T = (\pi_1, \mu_1, \mu_0, \eta_1, \eta_0)$, $\alpha^T = (\beta_0, \beta, \gamma)$, we can obtain that
$\sqrt{n}(\hat{\pi}_1 - \pi_1) \sim \mathcal{AN}(0, \pi_1\pi_0)$,
$\sqrt{n}(\hat{\mu}_1 - \mu_1) \sim \mathcal{AN}(0, \frac{1}{\pi_1\eta_1} = \frac{1}{\pi_1})$, $\sqrt{n}(\hat{\mu}_0 - \mu_0) \sim \mathcal{AN}(0, \frac{1}{\pi_0\eta_0} = \frac{\rho}{\pi_0})$,
$\sqrt{n}(\hat{\eta}_1 - \eta_1) \sim \mathcal{AN}(0, \frac{2\eta_1^2}{\pi_1} = \frac{2}{\pi_1})$, $\sqrt{n}(\hat{\eta}_0 - \eta_0) \sim \mathcal{AN}(0, \frac{2\eta_0^2}{\pi_0} = \frac{2}{\pi_0\rho^2})$.
It then follows that

$$\Sigma_{\mathrm{g}}(\hat{\theta}) = \mathrm{Diag}(\Sigma_{\mathrm{g}}(\hat{\pi}_1), \Sigma_{\mathrm{g}}(\hat{\mu}_1), \Sigma_{\mathrm{g}}(\hat{\mu}_0), \Sigma_{\mathrm{g}}(\hat{\eta}_1), \Sigma_{\mathrm{g}}(\hat{\eta}_0)) = \mathrm{Diag}(\pi_1\pi_0, \frac{1}{\pi_1}, \frac{\rho}{\pi_0}, \frac{2}{\pi_1}, \frac{2}{\pi_0\rho^2}) ,$$

and

$$\Sigma_{\rm g}(\hat{\alpha}) = M\Sigma_{\rm g}(\hat{\theta})M^T = \frac{1}{\pi_1\pi_0}\begin{bmatrix} \frac{\pi_0\mu^4+3}{2} & -\pi_0\mu^3 & \frac{\pi_0\mu^2-\pi_0-\frac{\pi_1}{\rho}}{2} \\ -\pi_0\mu^3 & \frac{\rho\pi_0(1+2\mu^2)+\pi_1}{\rho} & -\pi_0\mu \\ \frac{\pi_0\mu^2-\pi_0-\frac{\pi_1}{\rho}}{2} & -\pi_0\mu & \frac{\pi_0+\frac{\pi_1}{\rho^2}}{2} \end{bmatrix}.$$

We note that our results for $[\Sigma_{\rm g}(\hat{\alpha})]_{1,1}$, $[\Sigma_{\rm g}(\hat{\alpha})]_{1,3}$, $[\Sigma_{\rm g}(\hat{\alpha})]_{3,1}$ and $[\Sigma_{\rm g}(\hat{\alpha})]_{3,3}$ are different from those in O'Neill (1980), which appear to be contain minor errors.

A.2.7.2 Re-calculation of $I_{y|\mathbf{x}}(\theta)$, $\Sigma_\lambda(\hat{\theta})$ and $\Sigma_\lambda(\hat{\alpha})$

After some algebra, we can obtain that

$\frac{1}{n}I_{y|\mathbf{x}}(\pi_1) = \frac{H_0}{\pi_1\pi_0}$,

$\frac{1}{n}I_{y|\mathbf{x}}(\mu_1) = \pi_1\pi_0(H_2 - 2\mu H_1 + \mu^2 H_0)$, $\frac{1}{n}I_{y|\mathbf{x}}(\mu_0) = \pi_1\pi_0\frac{H_2}{\rho^2}$,

$\frac{1}{n}I_{y|\mathbf{x}}(\eta_1) = \frac{\pi_1\pi_0}{4}(H_4 - 4\mu H_3 + (6\mu^2-2)H_2 - 4\mu(\mu^2-1)H_1 + (\mu^2-1)^2 H_0)$,

$\frac{1}{n}I_{y|\mathbf{x}}(\eta_0) = \frac{\pi_1\pi_0}{4}(H_4 - 2\rho H_2 + \rho^2 H_0)$,

where $H_i = \int_{-\infty}^{\infty}\frac{p(x|\theta_1)p(x|\theta_0)x^i}{p(x)}dx,\ i = 0, 1, \ldots$. More precisely, H_i can be evaluated numerically as

$$H_i = \int_{-\infty}^{\infty}\frac{\frac{1}{\sqrt{2\pi\rho}}e^{-\frac{x^2}{2\rho}}x^i}{\pi_1 + \pi_0\frac{1}{\sqrt{\rho}}e^{\frac{(x-\mu)^2\rho - x^2}{2\rho}}}dx,\ i = 0, 1, \ldots.$$

Lemma A.2.1 *$H_{2k} \geq 0$, $k = 0, 1, \ldots$, and H_{2k} is even-symmetric whereas H_{2k+1} is odd-symmetric about $\mu = 0$.* ∎

As with Section A.1.7.2, using $I_{y|\mathbf{x}}(\theta)$ and $I_{\rm g}(\theta)$ (Section A.2.7.1), we can first derive $I_\lambda(\theta)$ and $V_\lambda(\theta)$ and then derive $\Sigma_\lambda(\hat{\theta})$ and $\Sigma_\lambda(\hat{\alpha})$, leading to

$$\Sigma_\lambda(\hat{\theta}) = \text{Diag}(\Sigma_\lambda(\hat{\pi}_1), \Sigma_\lambda(\hat{\mu}_1), \Sigma_\lambda(\hat{\mu}_0), \Sigma_\lambda(\hat{\eta}_1), \Sigma_\lambda(\hat{\eta}_0)),$$

where

$$\Sigma_\lambda(\hat{\pi}_1) = \pi_1\pi_0\frac{\lambda^2 + (1-\lambda^2)H_0}{[\lambda + (1-\lambda)H_0]^2},$$

$$\Sigma_\lambda(\hat{\mu}_1) = \frac{1}{\pi_1}\frac{\lambda^2 + (1-\lambda^2)\pi_0(H_2 - 2\mu H_1 + \mu^2 H_0)}{[\lambda + (1-\lambda)\pi_0(H_2 - 2\mu H_1 + \mu^2 H_0)]^2},$$

$$\Sigma_\lambda(\hat{\mu}_0) = \frac{1}{\pi_0}\frac{\lambda^2\frac{1}{\rho} + (1-\lambda^2)\pi_1\frac{H_2}{\rho^2}}{[\lambda\frac{1}{\rho} + (1-\lambda)\pi_1\frac{H_2}{\rho^2}]^2},$$

$$\Sigma_\lambda(\hat{\eta}_1) = \frac{\frac{\pi_1}{2}\lambda^2 + (1-\lambda^2)\frac{\pi_0\pi_1}{4}(H_4 - 4\mu H_3 + (6\mu^2-2)H_2 - 4\mu(\mu^2-1)H_1 + (\mu^2-1)^2 H_0)}{[\frac{\pi_1}{2}\lambda + (1-\lambda)\frac{\pi_0\pi_1}{4}(H_4 - 4\mu H_3 + (6\mu^2-2)H_2 - 4\mu(\mu^2-1)H_1 + (\mu^2-1)^2 H_0)]^2},$$

and

$$\Sigma_\lambda(\hat{\eta}_0) = \frac{\frac{\pi_0\rho^2}{2}\lambda^2 + (1-\lambda^2)\frac{\pi_0\pi_1}{4}(H_4 - 2\rho H_2 + \rho^2 H_0)}{[\frac{\pi_0\rho^2}{2}\lambda + (1-\lambda)\frac{\pi_0\pi_1}{4}(H_4 - 2\rho H_2 + \rho^2 H_0)]^2} .$$

Thus, we obtain

$$\Sigma_\lambda(\hat{\alpha}) = \begin{bmatrix} [\Sigma_\lambda(\hat{\alpha})]_{1,1} & [\Sigma_\lambda(\hat{\alpha})]_{1,2} & [\Sigma_\lambda(\hat{\alpha})]_{1,3} \\ [\Sigma_\lambda(\hat{\alpha})]_{2,1} & [\Sigma_\lambda(\hat{\alpha})]_{2,2} & [\Sigma_\lambda(\hat{\alpha})]_{2,3} \\ [\Sigma_\lambda(\hat{\alpha})]_{3,1} & [\Sigma_\lambda(\hat{\alpha})]_{3,2} & [\Sigma_\lambda(\hat{\alpha})]_{3,3} \end{bmatrix} ,$$

where

$$[\Sigma_\lambda(\hat{\alpha})]_{1,1} = (\frac{1}{\pi_0\pi_1})^2\Sigma_\lambda(\hat{\pi}_1) + \mu^2\Sigma_\lambda(\hat{\mu}_1) + [\frac{1-\mu^2}{2}]^2\Sigma_\lambda(\hat{\eta}_1) + (\frac{\rho}{2})^2\Sigma_\lambda(\hat{\eta}_0) ,$$

$$[\Sigma_\lambda(\hat{\alpha})]_{1,2} = [\Sigma_\lambda(\hat{\alpha})]_{2,1} = -\mu\Sigma_\lambda(\hat{\mu}_1) + \frac{\mu(1-\mu^2)}{2}\Sigma_\lambda(\hat{\eta}_1) ,$$

$$[\Sigma_\lambda(\hat{\alpha})]_{1,3} = [\Sigma_\lambda(\hat{\alpha})]_{3,1} = -\frac{1-\mu^2}{4}\Sigma_\lambda(\hat{\eta}_1) - \frac{\rho}{4}\Sigma_\lambda(\hat{\eta}_0) ,$$

$$[\Sigma_\lambda(\hat{\alpha})]_{2,2} = \Sigma_\lambda(\hat{\mu}_1) + (\frac{1}{\rho})^2\Sigma_\lambda(\hat{\mu}_0) + \mu^2\Sigma_\lambda(\hat{\eta}_1) ,$$

$$[\Sigma_\lambda(\hat{\alpha})]_{2,3} = [\Sigma_\lambda(\hat{\alpha})]_{3,2} = -\frac{\mu}{2}\Sigma_\lambda(\hat{\eta}_1) ,$$

$$[\Sigma_\lambda(\hat{\alpha})]_{3,3} = \frac{1}{4}\Sigma_\lambda(\hat{\eta}_1) + \frac{1}{4}\Sigma_\lambda(\hat{\eta}_0) .$$

A.2.7.3 Re-calculation of $\Sigma_{\mathrm{d}}(\hat{\alpha})$

Since $\nabla_\alpha g(x,\alpha) = \begin{pmatrix} 1 & x & x^2 \end{pmatrix}^T$, after some algebra, we obtain

$$\Sigma_{\mathrm{d}}^{-1}(\hat{\alpha}) = \pi_1\pi_0 \begin{bmatrix} H_0 & H_1 & H_2 \\ H_1 & H_2 & H_3 \\ H_2 & H_3 & H_4 \end{bmatrix} \triangleq \pi_1\pi_0\mathcal{W} ,$$

where H_i is as defined earlier in Section A.2.7.2. It follows that

$$\Sigma_{\mathrm{d}}(\hat{\alpha}) = \frac{1}{\pi_1\pi_0 \det(\mathcal{W})} \begin{bmatrix} H_2H_4 - H_3^2 & H_2H_3 - H_1H_4 & H_1H_3 - H_2^2 \\ H_2H_3 - H_1H_4 & H_0H_4 - H_2^2 & H_1H_2 - H_0H_3 \\ H_1H_3 - H_2^2 & H_1H_2 - H_0H_3 & H_0H_2 - H_1^2 \end{bmatrix} ,$$

where $\det(\mathcal{W}) = H_0(H_2H_4 - H_3^2) + H_1(H_2H_3 - H_1H_4) + H_2(H_1H_3 - H_2^2)$.

A.2.7.4 Re-calculation of B

For $g(x,\alpha) = \beta_0 + \beta x + \gamma x^2$, $x|\theta_1 \sim \mathcal{N}(\mu, 1)$, $x|\theta_0 \sim \mathcal{N}(0, \rho)$, where $\rho \in (0,1) \cup (1,\infty)$. After some algebra, we obtain $\beta_0 = \log\frac{\pi_1}{\pi_0} + \frac{\log\rho - \mu^2}{2}$, $\beta = \mu$, $\gamma = \frac{1-\rho}{2\rho}$, so $|\nabla_x g(x,\alpha)| = |\beta + 2\gamma x|$, $\nabla_\alpha g(x,\alpha) = (1\ x\ x^2)^T$, and
$g(\tilde{x},\alpha) = 0 \Leftrightarrow \tilde{x} = \frac{-\beta \pm \sqrt{\beta^2 - 4\beta_0\gamma}}{2\gamma}$.

Let $\tilde{\Delta} = \sqrt{\beta^2 - 4\beta_0\gamma}$, $\tilde{x}_1 = \frac{-\beta+\tilde{\Delta}}{2\gamma}$, $\tilde{x}_2 = \frac{-\beta-\tilde{\Delta}}{2\gamma}$. Then, given $\tilde{\Delta} \in [0,\infty)$, since

$$\nabla_x g(x,\alpha)|_{x=\tilde{x}_i} = \beta + 2\gamma\tilde{x}_i = \pm\tilde{\Delta},\ p(\tilde{x}_i|\theta_1) = \phi(\tilde{x}_i - \mu) = \phi\left(\frac{-\beta \pm \rho\tilde{\Delta}}{1-\rho}\right), i = 1,2,$$

O'Neill (1980) showed that

$$\mathbf{B} = \frac{\pi_1}{2\tilde{\Delta}}\left[\begin{pmatrix}1\\ \tilde{x}_2\\ \tilde{x}_2^2\end{pmatrix}\begin{pmatrix}1\\ \tilde{x}_2\\ \tilde{x}_2^2\end{pmatrix}^T \phi\left(\frac{-\beta-\rho\tilde{\Delta}}{1-\rho}\right) + \begin{pmatrix}1\\ \tilde{x}_1\\ \tilde{x}_1^2\end{pmatrix}\begin{pmatrix}1\\ \tilde{x}_1\\ \tilde{x}_1^2\end{pmatrix}^T \phi\left(\frac{-\beta+\rho\tilde{\Delta}}{1-\rho}\right)\right].$$

A.2.8 Numerical Evaluations of ARE

We can represent $\Sigma_{\mathrm{g}}(\hat{\alpha})$, $\Sigma_{\lambda}(\hat{\alpha})$ and $\Sigma_{\mathrm{d}}(\hat{\alpha})$ in a general notation as

$$\Sigma(\hat{\alpha}) = \begin{bmatrix}\Sigma_{1,1}^{(\hat{\alpha})} & \Sigma_{1,2}^{(\hat{\alpha})} & \Sigma_{1,3}^{(\hat{\alpha})}\\ \Sigma_{1,2}^{(\hat{\alpha})} & \Sigma_{2,2}^{(\hat{\alpha})} & \Sigma_{2,3}^{(\hat{\alpha})}\\ \Sigma_{1,3}^{(\hat{\alpha})} & \Sigma_{2,3}^{(\hat{\alpha})} & \Sigma_{3,3}^{(\hat{\alpha})}\end{bmatrix}.$$

Along with $\mathbf{B}$ as derived in Section A.2.7.4, it follows that

$$\mathrm{tr}(\mathbf{B}\Sigma(\hat{\alpha})) = \frac{\pi_1}{2\tilde{\Delta}}\left[\zeta_2^{(\hat{\alpha})}\phi\left(\frac{-\beta-\rho\tilde{\Delta}}{1-\rho}\right) + \zeta_1^{(\hat{\alpha})}\phi\left(\frac{-\beta+\rho\tilde{\Delta}}{1-\rho}\right)\right],$$

where $\zeta_1^{(\hat{\alpha})} = \Sigma_{1,1}^{(\hat{\alpha})} + 2\tilde{x}_1\Sigma_{1,2}^{(\hat{\alpha})} + \tilde{x}_1^2\left(2\Sigma_{1,3}^{(\hat{\alpha})} + \Sigma_{2,2}^{(\hat{\alpha})}\right) + 2\tilde{x}_1^3\Sigma_{2,3}^{(\hat{\alpha})} + \tilde{x}_1^4\Sigma_{3,3}^{(\hat{\alpha})}$ and $\zeta_2^{(\hat{\alpha})} = \Sigma_{1,1}^{(\hat{\alpha})} + 2\tilde{x}_2\Sigma_{1,2}^{(\hat{\alpha})} + \tilde{x}_2^2\left(2\Sigma_{1,3}^{(\hat{\alpha})} + \Sigma_{2,2}^{(\hat{\alpha})}\right) + 2\tilde{x}_2^3\Sigma_{2,3}^{(\hat{\alpha})} + \tilde{x}_2^4\Sigma_{3,3}^{(\hat{\alpha})}$. Therefore,

$$\mathrm{ARE}(\hat{\alpha}_2,\hat{\alpha}_1) = \frac{\mathrm{tr}\left(\mathbf{B}\Sigma(\hat{\alpha}_1)\right)}{\mathrm{tr}\left(\mathbf{B}\Sigma(\hat{\alpha}_2)\right)} = \frac{\zeta_2^{(\hat{\alpha}_1)}\phi\left(\frac{-\beta-\rho\tilde{\Delta}}{1-\rho}\right) + \zeta_1^{(\hat{\alpha}_1)}\phi\left(\frac{-\beta+\rho\tilde{\Delta}}{1-\rho}\right)}{\zeta_2^{(\hat{\alpha}_2)}\phi\left(\frac{-\beta-\rho\tilde{\Delta}}{1-\rho}\right) + \zeta_1^{(\hat{\alpha}_2)}\phi\left(\frac{-\beta+\rho\tilde{\Delta}}{1-\rho}\right)}.$$

Numerical evaluations of the ARE between the the generative, discriminative and GDT approaches for the quadratic normal discrimination are carried out under the conditions (1) $x|\theta_1 \sim \mathcal{N}(\mu,1)$, $x|\theta_0 \sim \mathcal{N}(0,\rho)$, (2) $\pi_1 = 0.5$, (3) $\rho \in [0.1, 2.0]$, (4) $\mu \in [-5,5]$ and (5) $\lambda \in [0,1]$.

A.2.8.1 Discriminative vs. Generative

Substituting $\Sigma_{\mathrm{g}}(\hat{\alpha})$ for $\Sigma(\hat{\alpha}_1)$ in $\zeta_1^{(\hat{\alpha}_1)}$ and $\zeta_2^{(\hat{\alpha}_1)}$, and substituting $\Sigma_{\mathrm{d}}(\hat{\alpha})$ for $\Sigma(\hat{\alpha}_2)$ in $\zeta_1^{(\hat{\alpha}_2)}$ and $\zeta_2^{(\hat{\alpha}_2)}$, we have $\mathrm{ARE}(\hat{\alpha}_{\mathrm{d}}, \hat{\alpha}_{\mathrm{g}})$ and denote it hereafter by $\mathrm{qEff}_{p=1}$.

Lemma A.2.2 *$qEff_{p=1}$ is even-symmetric about $\mu = 0$.* ∎

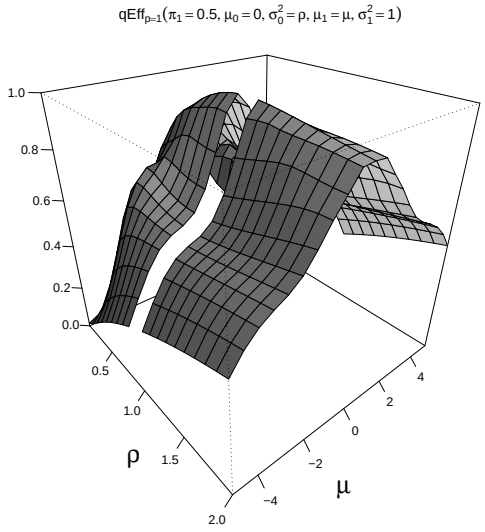


Figure A.1: The ARE between the generative approach and the discriminative approach for quadratic normal discrimination: $\mathrm{qEff}_{p=1}$ is the ARE for one-dimensional data. In the plot the gap is for $\rho = 1$ where the quadratic discrimination degenerates into a linear one.

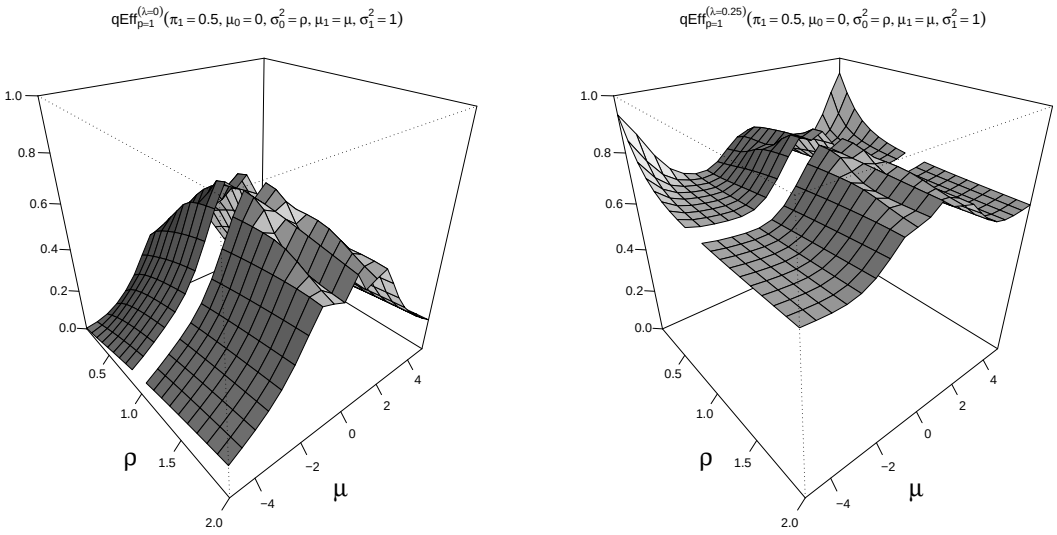
The numerical evaluation of $\mathrm{qEff}_{p=1}$ is shown in Figure A.1; we can make similar observations about $\mathrm{qEff}_{p=1}$ to those we made about $\mathrm{Eff}_{p=1}$ in Section 3.2.3.1.

A.2.8.2 Trade-off vs. Generative

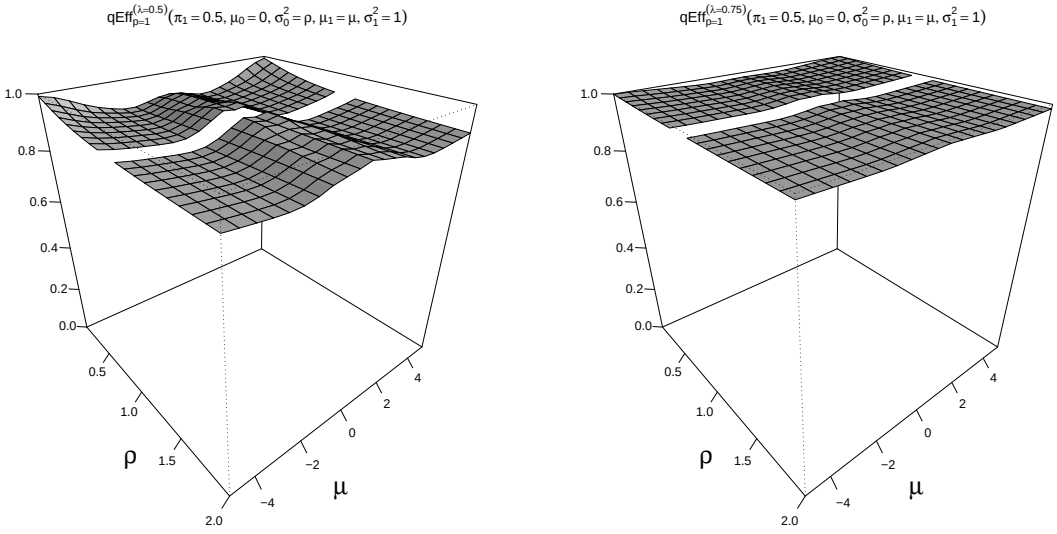
Substituting $\Sigma_{\mathrm{g}}(\hat{\alpha})$ for $\Sigma(\hat{\alpha}_1)$ in $\zeta_1^{(\hat{\alpha}_1)}$ and $\zeta_2^{(\hat{\alpha}_1)}$, and substituting $\Sigma_{\lambda}(\hat{\alpha})$ for $\Sigma(\hat{\alpha}_2)$ in $\zeta_1^{(\hat{\alpha}_2)}$ and $\zeta_2^{(\hat{\alpha}_2)}$, we have $\mathrm{ARE}(\hat{\alpha}_{\lambda}, \hat{\alpha}_{\mathrm{g}})$ and denote it hereafter by $\mathrm{qEff}_{p=1}^{(\lambda)}$.

Lemma A.2.3 *$qEff_{p=1}^{(\lambda)}$ is even-symmetric about $\mu = 0$.* ∎

Numerical evaluations of $\mathrm{qEff}_{p=1}^{(\lambda)}$, with $\lambda = 0, 0.25, 0.5$ and 0.75 respectively, are shown in Figure A.2; we can make similar observations about $\mathrm{qEff}_{p=1}^{(\lambda)}$ to those we made about $\mathrm{Eff}_{p=1}^{(\lambda)}$ in Section 3.2.3.2.



(a) qEff$_{p=1}^{(\lambda=0)}$ (b) qEff$_{p=1}^{(\lambda=0.25)}$



(c) qEff$_{p=1}^{(\lambda=0.5)}$ (d) qEff$_{p=1}^{(\lambda=0.75)}$

Figure A.2: The ARE between the generative approach and the GDT with $\lambda = 0, 0.25, 0.5$ and 0.75 respectively, for quadratic normal discrimination.

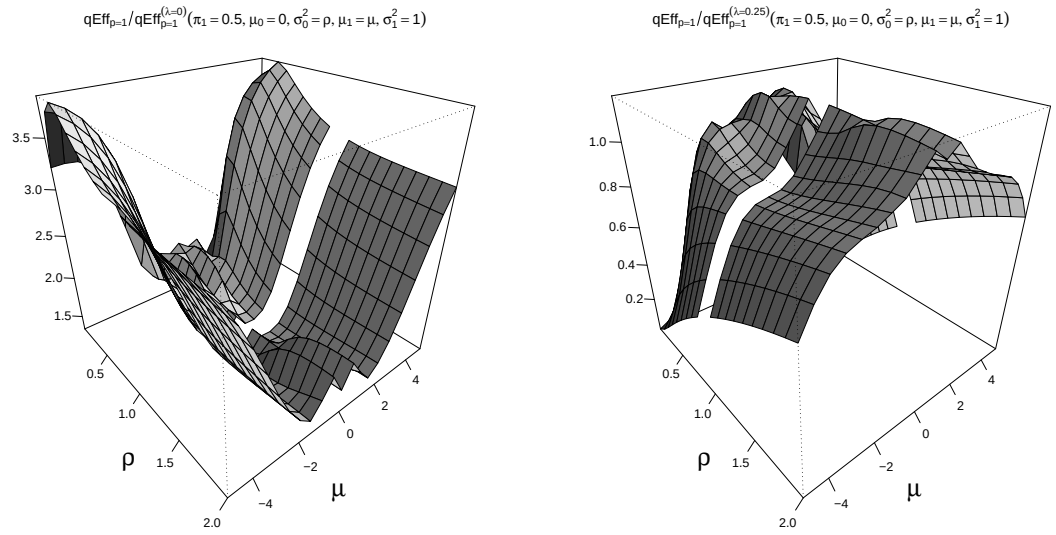
A.2.8.3 Discriminative vs. Trade-off

The ARE between the discriminative approach and the GDT is simply

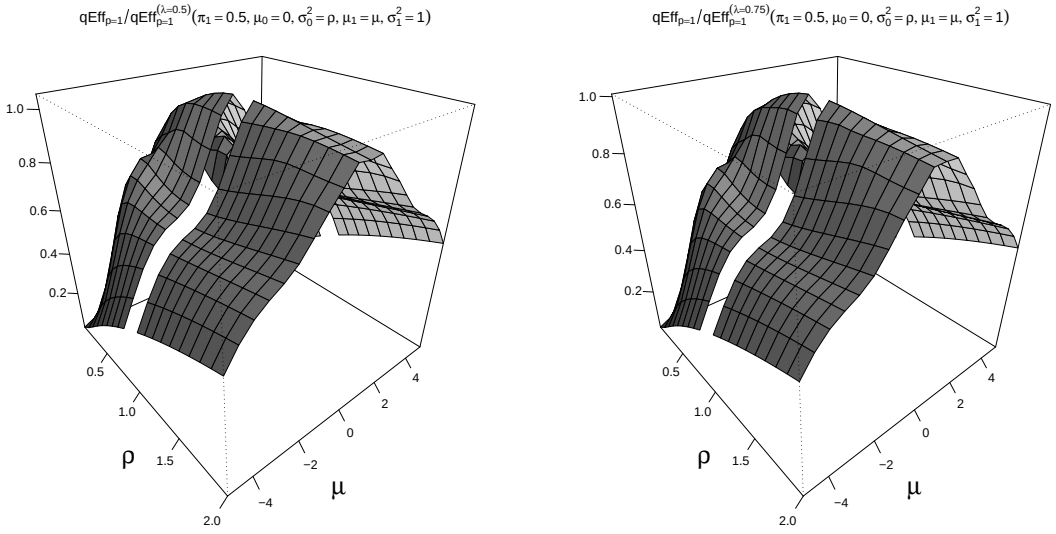
$$\text{ARE}(\hat{\alpha}_{\text{d}}, \hat{\alpha}_{\lambda}) = \frac{\text{ARE}(\hat{\alpha}_{\text{d}}, \hat{\alpha}_{\text{g}})}{\text{ARE}(\hat{\alpha}_{\lambda}, \hat{\alpha}_{\text{g}})} = \frac{\text{qEff}_{p=1}}{\text{qEff}_{p=1}^{(\lambda)}} \, .$$

Lemma A.2.4 *ARE*$(\hat{\alpha}_d, \hat{\alpha}_\lambda)$ *is even-symmetric about* $\mu = 0$. ∎

Numerical evaluations of $\frac{\text{qEff}_{p=1}}{\text{qEff}_{p=1}^{(\lambda)}}$, for $\lambda = 0, 0.25, 0.5$ and 0.75, respectively, are shown in Figure A.3; we can make similar observations about $\frac{\text{qEff}_{p=1}}{\text{qEff}_{p=1}^{(\lambda)}}$ to those we made about $\frac{\text{Eff}_{p=1}}{\text{Eff}_{p=1}^{(\lambda)}}$ in Section 3.2.3.3.



(a) qEff$_{p=1}$/qEff$_{p=1}^{(\lambda=0)}$ (b) qEff$_{p=1}$/qEff$_{p=1}^{(\lambda=0.25)}$



(c) qEff$_{p=1}$/qEff$_{p=1}^{(\lambda=0.5)}$ (d) qEff$_{p=1}$/qEff$_{p=1}^{(\lambda=0.75)}$

Figure A.3: The ARE between the GDT and the discriminative approach with $\lambda = 0, 0.25, 0.5$ and 0.75 respectively, for quadratic normal discrimination.

Appendix B

Appendix for Chapter 4

B.1 Results for Simulated Discrete Data

B.1.1 With a Common Covariance Matrix Σ

The third set of 3 datasets contains simulated discrete data arising from two 4-variate Bernoulli distributions: $\mathbf{x} \sim B(\mathbf{p})$ for the group with $y = 1$ and $\mathbf{x} \sim B(\mathbf{q})$ for $y = 2$, where $\mathbf{p} = (p_1, p_2, p_3, p_4)^T = (0.2, 0.3, 0.4, 0.5)^T$, $\mathbf{q} = (q_1, q_2, q_3, q_4)^T = (0.8, 0.7, 0.6, 0.5)^T$. In this context, the two groups have a common covariance matrix Σ but different means ($\mu_1 = E\{\mathbf{x}|y = 1\} = \mathbf{p}$ and $\mu_2 = E\{\mathbf{x}|y = 2\} = \mathbf{q}$). Σ is a diagonal, block diagonal and full covariance matrix, respectively for these 3 datasets.

For the first dataset, each of the 4 features $\{x_j\}_{j=1}^4$ is conditionally independent of the others given the group label y. In order to achieve this, we set all the elements of $\mathbf{p}$ and $\mathbf{q}$ such that the covariance matrices for the two groups are diagonal matrices:

$$\Sigma_{y=1} = \text{diag}(V_{1,1}, V_{2,2}, V_{3,3}, V_{4,4}) \ , \ \Sigma_{y=2} = \text{diag}(V'_{1,1}, V'_{2,2}, V'_{3,3}, V'_{4,4}) \ ,$$

where, for $i = 1, \ldots, 4$,

$$V_{i,i} = p_i \left(1 - p_i\right) \ , \ V'_{i,i} = q_i \left(1 - q_i\right) \ .$$

In order to have $\Sigma_{y=1} = \Sigma_{y=2} = \Sigma$, we set $q_i = 1 - p_i$.

For the second dataset, $\mathbf{x}^1$ is conditionally independent of $\mathbf{x}^2$ given the group label y. In order to achieve this, we set only p_1, p_3, q_1, q_3 and conditional probabilities $p_{2|1(1)}, p_{2|1(0)}, p_{4|3(1)}, p_{4|3(0)}$ and $q_{2|1(1)}, q_{2|1(0)}, q_{4|3(1)}, q_{4|3(0)}$, where $p_{i|j(v)}$ and $q_{i|j(v)}$ denote the success probabilities p_i

and q_i of x_i given $x_j = v, v \in 0, 1$. It follows that

$$p_2 = p_1 p_{2|1(1)} + (1 - p_1) p_{2|1(0)} \ , \ q_2 = q_1 q_{2|1(1)} + (1 - q_1) q_{2|1(0)} \ ,$$

$$p_4 = p_3 p_{4|3(1)} + (1 - p_3) p_{4|3(0)} \ , \ q_4 = q_3 q_{4|3(1)} + (1 - q_3) q_{4|3(0)} \ ,$$

and the covariance matrices for the two groups are block diagonal, symmetric matrices:

$$\Sigma_{y=1} = \begin{bmatrix} V_{1,1} & V_{1,2} & 0 & 0 \\ V_{1,2} & V_{2,2} & 0 & 0 \\ 0 & 0 & V_{3,3} & V_{3,4} \\ 0 & 0 & V_{3,4} & V_{4,4} \end{bmatrix} \ , \ \Sigma_{y=2} = \begin{bmatrix} V'_{1,1} & V'_{1,2} & 0 & 0 \\ V'_{1,2} & V'_{2,2} & 0 & 0 \\ 0 & 0 & V'_{3,3} & V'_{3,4} \\ 0 & 0 & V'_{3,4} & V'_{4,4} \end{bmatrix} \ ,$$

where, for $i = 1, \ldots, 4$,

$$V_{i,i} = p_i \left(1 - p_i\right) \ , \ V'_{i,i} = q_i \left(1 - q_i\right) \ ,$$

$$V_{1,2} = p_1 \left(p_{2|1(1)} - p_2\right) \ , \ V'_{1,2} = q_1 \left(q_{2|1(1)} - q_2\right) \ ,$$

$$V_{3,4} = p_3 \left(p_{4|3(1)} - p_4\right) \ , \ V'_{3,4} = q_3 \left(q_{4|3(1)} - q_4\right) \ .$$

In order to have $\Sigma_{y=1} = \Sigma_{y=2} = \Sigma$, we set

$$q_1 = 1 - p_1 \ , \ q_3 = 1 - p_3 \ ,$$

$$q_{2|1(1)} = 1 - p_{2|1(0)} \ , \ q_{2|1(0)} = 1 - p_{2|1(1)} \ , \text{ and}$$

$$q_{4|3(1)} = 1 - p_{4|3(0)} \ , \ q_{4|3(0)} = 1 - p_{4|3(1)} \ .$$

For the third dataset, each of the 4 features $\{x_j\}_{j=1}^4$ is dependent on the others given the group label y. In order to achieve that, we set only p_1, q_1 and conditional probabilities $p_{i|1(1)}, p_{i|1(0)}$ and $q_{i|1(1)}, q_{i|1(0)}$, for $i = 2, 3, 4$. It follows that, for $i = 2, 3, 4$,

$$p_i = p_1 p_{i|1(1)} + (1 - p_1) p_{i|1(0)} \ , \ q_i = q_1 q_{i|1(1)} + (1 - q_1) q_{i|1(0)} \ ,$$

and the covariance matrices for the two groups are full symmetric matrices:

$$\Sigma_{y=1} = \begin{bmatrix} V_{1,1} & V_{1,2} & V_{1,3} & V_{1,4} \\ V_{1,2} & V_{2,2} & V_{2,3} & V_{2,4} \\ V_{1,3} & V_{2,3} & V_{3,3} & V_{3,4} \\ V_{1,4} & V_{2,4} & V_{3,4} & V_{4,4} \end{bmatrix} \ , \ \Sigma_{y=2} = \begin{bmatrix} V'_{1,1} & V'_{1,2} & V'_{1,3} & V'_{1,4} \\ V'_{1,2} & V'_{2,2} & V'_{2,3} & V'_{2,4} \\ V'_{1,3} & V'_{2,3} & V'_{3,3} & V'_{3,4} \\ V'_{1,4} & V'_{2,4} & V'_{3,4} & V'_{4,4} \end{bmatrix} \ ,$$

where

$$V_{i,i} = p_i\,(1-p_i)\ ,\ V'_{i,i} = q_i\,(1-q_i)\ ,\ i = 1,\ldots,4\ ;$$

$$V_{1,i} = p_1\left(p_{i|1(1)} - p_i\right)\ ,\ V'_{1,i} = q_1\left(q_{i|1(1)} - q_i\right)\ ,\ i = 2,3,4\ ;$$

and, for $i, j = 2, 3, 4$,

$$p(x_i = 1, x_j = 1) = p_1 p_{i|1(1)} p_{j|1(1)} + (1-p_1) p_{i|1(0)} p_{j|1(0)}\ ,$$

$$q(x_i = 1, x_j = 1) = q_1 q_{i|1(1)} q_{j|1(1)} + (1-q_1) q_{i|1(0)} q_{j|1(0)}\ ,$$

such that

$$V_{i,j} = p(x_i = 1, x_j = 1) - p_i p_j\ ,\ V'_{i,j} = q(x_i = 1, x_j = 1) - q_i q_j\ .$$

In order to have $\Sigma_{y=1} = \Sigma_{y=2} = \Sigma$, we set

$$q_1 = 1 - p_1\ ,$$

$$q_{i|1(1)} = 1 - p_{i|1(0)}\ ,\ \text{and}\ q_{i|1(0)} = 1 - p_{i|1(1)}\ ,\ i = 2,3,4\ .$$

B.1.1.1 Diagonal Covariance Matrix Σ

For the first dataset, we set $\mu_1 = \mathbf{p} = (0.2, 0.3, 0.4, 0.5)^T$, $\mu_2 = \mathbf{q} = \mathbf{1} - \mathbf{p} = (0.8, 0.7, 0.6, 0.5)^T$ such that the common covariance matrix Σ is a diagonal matrix, diag$(0.16, 0.21, 0.24, 0.25)$.

B.1.1.2 Block Diagonal Covariance Matrix Σ

For the second dataset, we set

$$p_1 = 0.2\ ,\ q_1 = 1 - p_1 = 0.8\ ,$$

$$p_3 = 0.4\ ,\ q_3 = 1 - p_3 = 0.6\ ;$$

$$p_{2|1(1)} = 0.7\ ,\ p_{2|1(0)} = 0.2\ ,$$

$$q_{2|1(1)} = 1 - p_{2|1(0)} = 0.8\ ,\ q_{2|1(0)} = 1 - p_{2|1(1)} = 0.3\ ;$$

$$p_{4|3(1)} = 0.8\ ,\ p_{4|3(0)} = 0.3\ ,$$

$$q_{4|3(1)} = 1 - p_{4|3(0)} = 0.7\ ,\ \text{and}\ q_{4|3(0)} = 1 - p_{4|3(1)} = 0.2\ .$$

It follows that $\mu_1 = \mathbf{p} = (0.2, 0.3, 0.4, 0.5)^T$, $\mu_2 = \mathbf{q} = \mathbf{1} - \mathbf{p} = (0.8, 0.7, 0.6, 0.5)^T$, and Σ is a block diagonal matrix $\begin{bmatrix} 0.16 & 0.08 & 0 & 0 \\ 0.08 & 0.21 & 0 & 0 \\ 0 & 0 & 0.24 & 0.12 \\ 0 & 0 & 0.12 & 0.25 \end{bmatrix}$.

B.1.1.3 Full Covariance Matrix Σ

For the third dataset, we set

$$p_1 = 0.2\ ,\ q_1 = 1 - p_1 = 0.8\ ;$$

$$p_{2|1(1)} = 0.7\ ,\ p_{2|1(0)} = 0.2\ ,$$

$$q_{2|1(1)} = 1 - p_{2|1(0)} = 0.8\ ,\ q_{2|1(0)} = 1 - p_{2|1(1)} = 0.3\ ;$$

$$p_{3|1(1)} = 0.8\ ,\ p_{3|1(0)} = 0.3\ ,$$

$$q_{3|1(1)} = 1 - p_{3|1(0)} = 0.7\ ,\ q_{3|1(0)} = 1 - p_{3|1(1)} = 0.2\ ;$$

$$p_{4|1(1)} = 0.9\ ,\ p_{4|1(0)} = 0.4\ ,$$

$$q_{4|1(1)} = 1 - p_{4|1(0)} = 0.6\ ,\ \text{and}\ q_{4|1(0)} = 1 - p_{4|1(1)} = 0.1\ .$$

It follows that $\mu_1 = \mathbf{p} = (0.2, 0.3, 0.4, 0.5)^T$, $\mu_2 = \mathbf{q} = \mathbf{1} - \mathbf{p} = (0.8, 0.7, 0.6, 0.5)^T$, and Σ is a full matrix $\begin{bmatrix} 0.16 & 0.08 & 0.08 & 0.08 \\ 0.08 & 0.21 & 0.04 & 0.04 \\ 0.08 & 0.04 & 0.24 & 0.04 \\ 0.08 & 0.04 & 0.04 & 0.25 \end{bmatrix}$.

The results for these 3 datasets are shown in Figure B.1.

B.1.2 With Unequal Covariance Matrices Σ_1, Σ_2

The settings of the last 3 datasets are similar to those of the third set in Section B.1.1, except that $\Sigma_1 \neq \Sigma_2$ and $\mathbf{q}$ is different amongst these 3 datasets.

B.1.2.1 Diagonal Covariance Matrices Σ_1, Σ_2

For the first dataset, the setting is the same as that in Section B.1.1.1 except that $\mathbf{q} = \mathbf{p} + 0.4$ rather than $\mathbf{q} = \mathbf{1} - \mathbf{p}$. That is, we set $\mu_1 = \mathbf{p} = (0.2, 0.3, 0.4, 0.5)^T$, $\mu_2 = \mathbf{q} = (0.6, 0.7, 0.8, 0.9)^T$ such that $\Sigma_1 = \text{diag}(0.16, 0.21, 0.24, 0.25)$ and $\Sigma_2 = \text{diag}(0.24, 0.21, 0.16, 0.09)$.

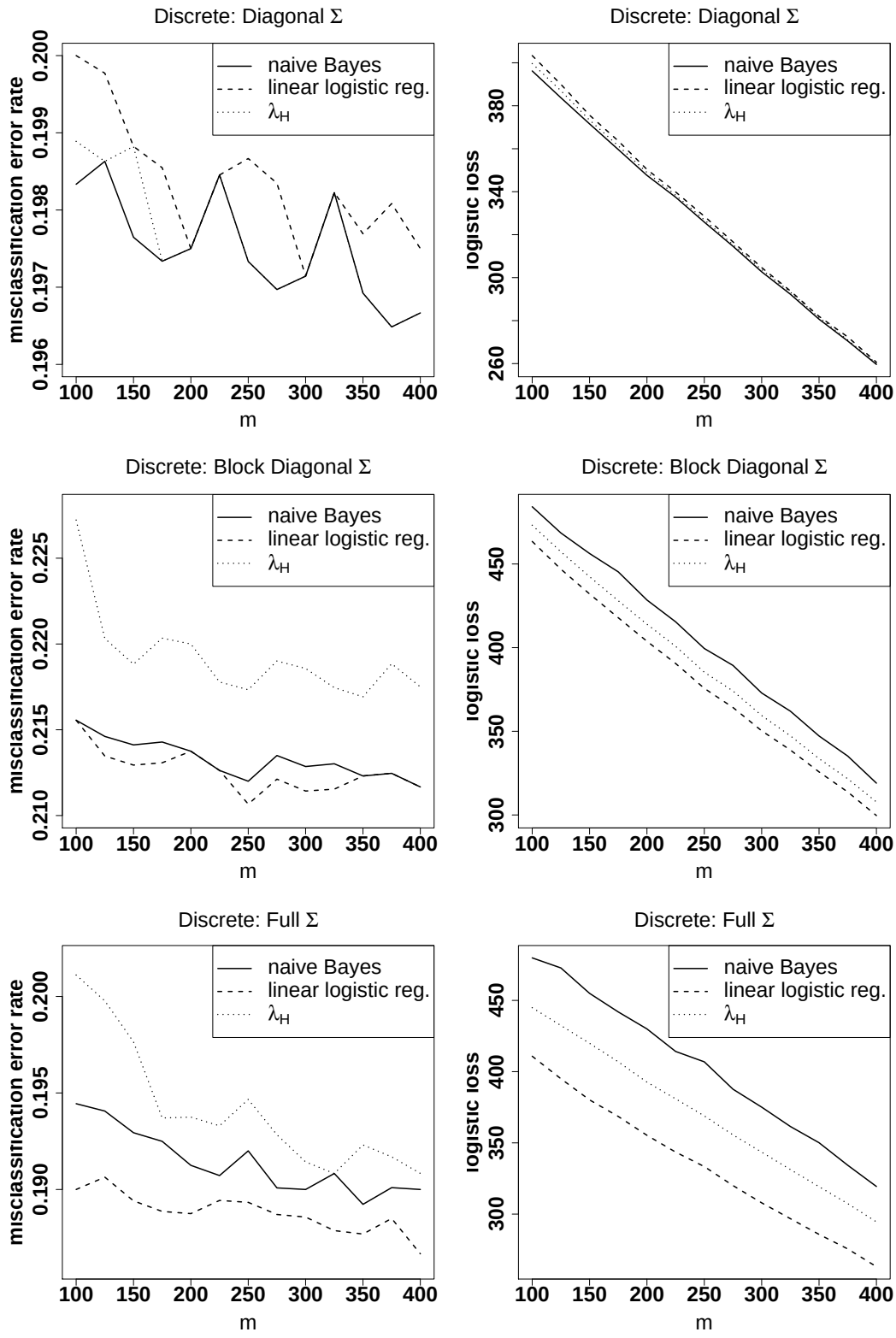


Figure B.1: Simulated Bernoulli data with equal covariance matrices. Plots of classification performance measured by ER and by LL vs. training set size m.

B.1.2.2 Block Diagonal Covariance Matrices Σ_1, Σ_2

For the second dataset, the setting is the same as that in Section B.1.1.2 except that $q_1 = p_1 + 0.4, q_3 = p_3 + 0.4$ rather than $q_1 = 1 - p_1, q_3 = 1 - p_3$, respectively. That is, we have $\mu_1 = \mathbf{p} = (0.2, 0.3, 0.4, 0.5)^T$, $\mu_2 = \mathbf{q} = (0.6, 0.6, 0.8, 0.6)^T$, $\Sigma_1 = \begin{bmatrix} 0.16 & 0.08 & 0 & 0 \\ 0.08 & 0.21 & 0 & 0 \\ 0 & 0 & 0.24 & 0.12 \\ 0 & 0 & 0.12 & 0.25 \end{bmatrix}$ and $\Sigma_2 = \begin{bmatrix} 0.24 & 0.12 & 0 & 0 \\ 0.12 & 0.24 & 0 & 0 \\ 0 & 0 & 0.16 & 0.08 \\ 0 & 0 & 0.08 & 0.24 \end{bmatrix}$.

B.1.2.3 Full Covariance Matrices Σ_1, Σ_2

For the third dataset, the setting is the same as that in Section B.1.1.3 except that $q_1 = p_1 + 0.4$ rather than $q_1 = 1 - p_1$. That is, we have $\mu_1 = \mathbf{p} = (0.2, 0.3, 0.4, 0.5)^T$, $\mu_2 = \mathbf{q} = (0.6, 0.6, 0.5, 0.4)^T$, $\Sigma_1 = \begin{bmatrix} 0.16 & 0.08 & 0.08 & 0.08 \\ 0.08 & 0.21 & 0.04 & 0.04 \\ 0.08 & 0.04 & 0.24 & 0.04 \\ 0.08 & 0.04 & 0.04 & 0.25 \end{bmatrix}$ and $\Sigma_2 = \begin{bmatrix} 0.24 & 0.12 & 0.12 & 0.12 \\ 0.12 & 0.24 & 0.06 & 0.06 \\ 0.12 & 0.06 & 0.25 & 0.06 \\ 0.12 & 0.06 & 0.06 & 0.24 \end{bmatrix}$; they are symmetric, positive-definite matrices.

The results for these 3 datasets are shown in Figure B.2.

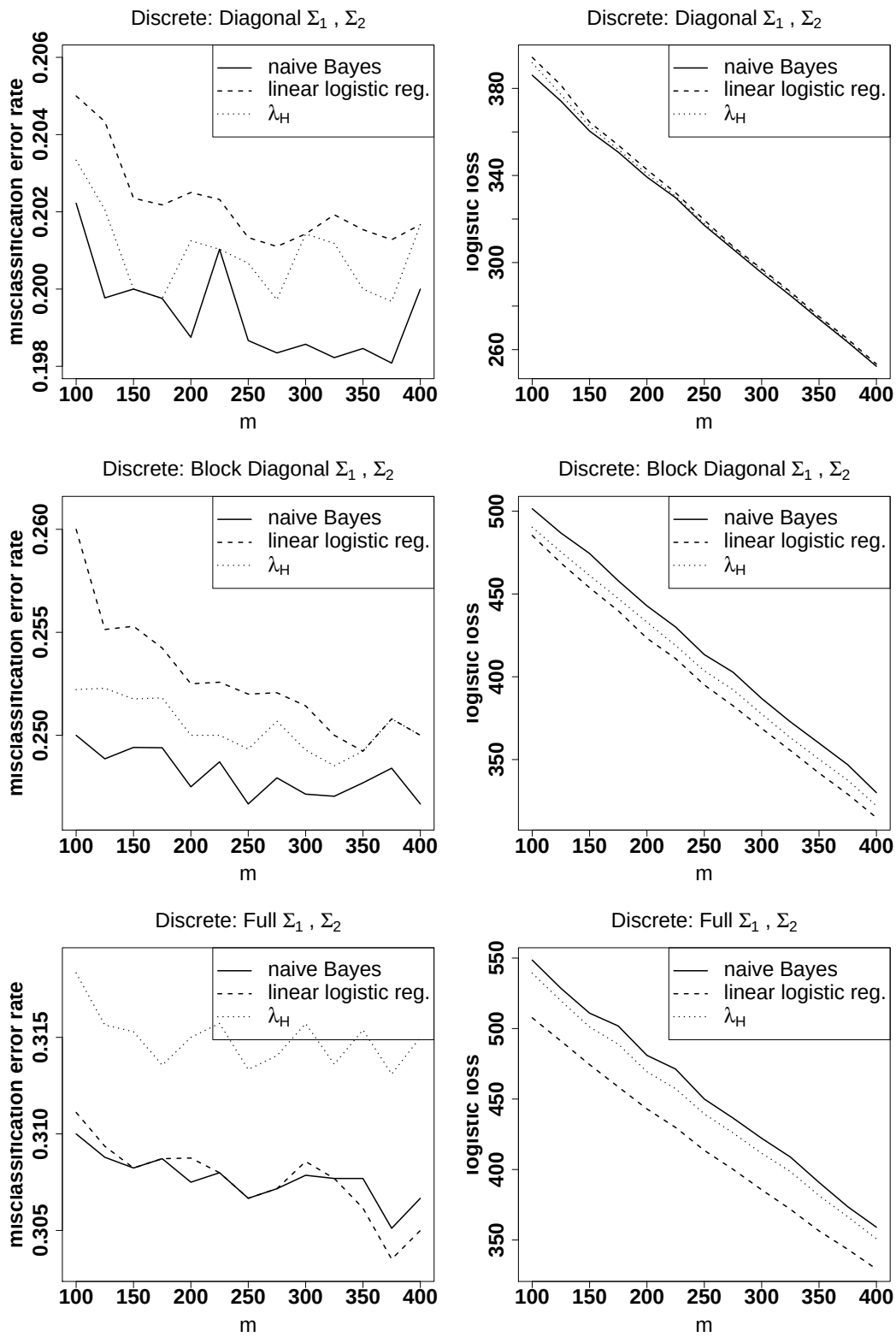


Figure B.2: Simulated Bernoulli data with unequal covariance matrices. Plots of classification performance measured by ER and by LL vs. training set size m.

Bibliography

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., Levine, A. J., 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. 96, 6745–6750.

Asuncion, A., Newman, D. J., 2007. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science, http://www.ics.uci.edu/∼mlearn/MLRepository.html.

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

Bishop, C. M., Lasserre, J., 2007. Generative or discriminative? Getting the best of both worlds (with discussion). In: Bayesian Statistics 8. pp. 3–24.

Bouchard, G., 2005. Generative models in supervised statistical learning with applications to digital image categorization and structural reliability. Ph.D. thesis, INRIA.

Bouchard, G., 2007. Bias-variance tradeoff in hybrid generative-discriminative models. In: ICMLA'07.

Bouchard, G., Triggs, B., 2004. The tradeoff between generative and discriminative classifiers. In: IASC International Symposium on Computational Statistics (COMPSTAT). pp. 721–728.

Dawid, A. P., 1976. Properties of diagnostic data distributions. Biometrics 32 (3), 647–658.

Druck, G., Pal, C., Zhu, X., McCallum, A., 2007. Semi-supervised classification with hybrid generative/discriminative methods. In: KDD. pp. 280–289.

Dudoit, S., Fridlyand, J., Speed, T. P., 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. J. Amer. Statistical Assoc. 97, 77–87.

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

Fan, J., Fan, Y., 2007. High dimensional classification using features annealed independence rules. Ann. Statist.To appear.

Fujino, A., Ueda, N., Saito, K., 2007. A hybrid generative/discriminative approach to text classification with additional information. Information Processing and Management 43 (2), 379–392.

Glasbey, C. A., 1993. An analysis of histogram-based thresholding algorithms. CVGIP: Graph. Models Image Process. 55 (6), 532–537.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., Lander, E. S., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–537.

Gonzalez, R. C., Woods, R. E., 2002. Digital Image Processing, 2nd Edition. Prentice Hall.

Hall, P., Titterington, D. M., Xue, J.-H., 2008. Tilting methods for assessing the influence of components in a classifier, manuscript.

Hand, D. J., 2006. Classifier technology and illusion of progress (with discussion). Statistical Science 21, 1–34.

Hand, D. J., Yu, K., 2001. Idiot's Bayes - not so stupid after all? International Statistical Review 69 (3), 385–398.

Kang, C., Tian, J., 2006. A hybrid generative/discriminative Bayesian classifier. In: FLAIRS Conference. pp. 562–567.

Kittler, J., Illingworth, J., 1986. Minimum error thresholding. Pattern Recogn. 19 (1), 41–47.

Krzanowski, W. J., 1983. Stepwise location model choice in mixed-variable discrimination. Applied Statistics 32 (3), 260–266.

Kurita, T., Otsu, N., Abdelmalek, N., 1992. Maximum likelihood thresholding based on population of mixture models. Pattern Recogn. 25 (10), 1231–1240.

Li, Y., 2005. Hidden Markov models with states depending on observations. Pattern Recognition Letters 26 (7), 977–984.

McCallum, A., Pal, C., Druck, G., Wang, X., 2006. Multi-conditional learning: Generative/discriminative training for clustering and classification. In: AAAI. pp. 433–439.

Moser, G., Serpico, S., 2006. Generalized minimum-error thresholding for unsupervised change detection from SAR amplitude imagery. IEEE Transactions on Geoscience and Remote Sensing 44 (10), 2972–2982.

Ng, A. Y., Jordan, M. I., 2001. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. In: NIPS. pp. 841–848.

Ni Bhrolchain, M., 1979. Psychotic and neurotic depression: I. Some points of method. British Journal of Psychiatry 134, 87–93.

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.

Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Trans. Systems Man Cybernet. SMC-9, 62–66.

Pal, N., Pal, S., 1993. A review on image segmentation techniques. Pattern Recogn. 26, 1277–1294.

Pal, N. R., Bhandari, D., 1993. Image thresholding: some new techniques. Signal Process. 33 (2), 139–158.

Perlich, C., Provost, F., Simonoff, J. S., 2003. Tree induction vs. logistic regression: A learning-curve analysis. Journal of Machine Learning Research 4 (211–255).

Raina, R., Shen, Y., Ng, A. Y., McCallum, A., 2003. Classification with hybrid generative/discriminative models. In: NIPS.

Rubinstein, Y. D., 1998. Discriminative vs informative learning. Ph.D. thesis, Stanford University.

Rubinstein, Y. D., Hastie, T., 1997. Discriminative vs. informative learning. In: KDD. pp. 49–53.

Sahoo, P. K., Soltani, S., Wong, A. K., Chen, Y. C., 1988. A survey of thresholding techniques. Comput. Vision Graph. Image Process. 41 (2), 233–260.

Sezgin, M., Sankur, B., 2004. Survey over image thresholding techniques and quantitative performance evaluation. Journal of Electronic Imaging 13 (1), 146–165.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., , Sellers, W. R., 2002. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1, 203–209.

Suzuki, J., Fujino, A., Isozaki, H., 2007. Semi-supervised structured output learning based on a hybrid generative and discriminative approach. In: EMNLP-CoNLL 2007. pp. 791–800.

Titterington, D. M., Murray, G. D., Murray, L. S., Spiegelhalter, D. J., Skene, A. M., Habbema, J. D. F., Gelpke, G. J., 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.

Trier, Ø. D., Jain, A. K., 1995. Goal-directed evaluation of binarization methods. IEEE Trans. Pattern Anal. Mach. Intell. 17 (12), 1191–1201.

Venables, W. N., Ripley, B. D., 2002. Modern Applied Statistics with S. Springer, New York.

Verboven, S., Hubert, M., 2005. LIBRA: a MATLAB library for robust analysis. Chemometrics and Intelligent Laboratory Systems 75 (2), 127–136.

Xue, J.-H., Zhang, Y. J., Lin, X. G., 1999. Rayleigh-distribution based minimum error thresholding for SAR images. Journal of Electronics (China) 16 (4), 336–342.

Yan, H., 1996. Unified formulation of a class of image thresholding techniques. Pattern Recogn. 29 (12), 2025–2032.

Zhang, H., Fritts, J. E., Goldman, S. A., 2007. Image segmentation evaluation: a survey of unsupervised methods. Comput. Vis. Image Understand. doi:10.1016/j.cviu.2007.08.003.

Zhang, Y. J., 1996. A survey on evaluation methods for image segmentation. Pattern Recogn. 29 (8), 1335–1346.

Zhou, G. D., 2005. Direct modelling of output context dependence in discriminative hidden Markov model. Pattern Recognition Letters 26 (5), 545–553.

List of Publications during PhD Study

1. J.-H. Xue and D. M. Titterington. 2007. Comment on "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes". (based on Chapter 2; submitted to *Neural Processing Letter*).

2. J.-H. Xue and D. M. Titterington. 2008. On the generative-discriminative tradeoff approach: Interpretation, asymptotic efficiency and classification performance. (based on Chapter 3; submitted to *Computational Statistics & Data Analysis*).

3. J.-H. Xue and D. M. Titterington. 2008. Interpretation of hybrid generative/discriminative algorithms. *Neurocomputing*. (based on Chapter 4; revised).

4. J.-H. Xue and D. M. Titterington. 2008. Joint generative-discriminative modelling based on statistical tests for classification. (based on Chapter 5; submitted to *Pattern Recognition Letters*).

5. J.-H. Xue and D. M. Titterington. 2008. Short note on two output-dependent hidden Markov models. *Pattern Recognition Letters*. (based on Chapter 6; doi:10.1016/j.patrec.2008.02.018).

6. J.-H. Xue and D. M. Titterington. 2007. Discriminative image thresholding. (based on Chapter 7; submitted to *Pattern Recognition*).

7. P. Hall, D. M. Titterington, and J.-H. Xue. 2008. Median-based classifiers for high-dimensional data. (submitted to *Journal of the American Statistical Association*).

8. P. Hall, D. M. Titterington, and J.-H. Xue. 2008. Tilting methods for assessing the influence of components in a classifier. (submitted to *Journal of the Royal Statistical Society: Series B*).

9. J.-H. Xue and D. M. Titterington. 2007. Short letter on the semiparametric transformation discriminant analysis. (to be submitted to *Biometrika*).

10. J.-H. Xue and D. M. Titterington. 2008. Do unbalanced data have a negative effect on LDA? *Pattern Recognition*, 41(5):1558–1571. (doi:10.1016/j.patcog.2007.11.008).