Elsevier Editorial System(tm) for Pattern Recognition

Manuscript Draft

Manuscript Number:

Title: Do Unbalanced Data Have a Negative Effect on LDA?

Article Type: Full Length Article

Keywords: Area under an ROC curve (AUC); Linear discriminant analysis (LDA); Misclassification error rate (ER); Unbalanced data

Corresponding Author: Mr. Jinghao Xue,

Corresponding Author's Institution:

First Author: Jinghao Xue

Order of Authors: Jinghao Xue; D. Michael Titterington

Manuscript Region of Origin:

Abstract: For two-class discrimination, Ref.~\cite{Xie:2007} claimed that, when covariance matrices of the two classes were unequal, a (class) unbalanced dataset had a negative effect on the performance of linear discriminant analysis (LDA). Through re-balancing 10 real-world datasets, Ref.~\cite{Xie:2007} provided empirical evidence to support the claim using AUC (Area Under the receiver operating characteristic Curve) as the performance metric. We suggest that such a claim is vague if not misleading, there is no solid theoretical analysis presented in~\cite{Xie:2007}, and AUC can lead to a quite different conclusion from that led to by misclassification error rate (ER) on the discrimination performance of LDA for unbalanced datasets. Our empirical and simulation studies

suggest that, for LDA, the increase of the median of AUC (and thus the improvement of performance of LDA) from re-balancing is relatively small, while, in contrast, the increase of the median of ER (and thus the decline in performance of LDA) from re-balancing is relatively large. Therefore, from our study, there is no reliable empirical evidence to support the claim that a (class) unbalanced data set has a negative effect on the performance of LDA. In addition, re-balancing affects the performance of LDA for datasets with either equal or unequal covariance matrices, indicating that having unequal covariance matrices is not a key reason for the difference in performance between original and re-balanced data.

Do Unbalanced Data Have a Negative Effect on LDA?

Jing-Hao Xue [a,*], D. Michael Titterington [a]

[a] *Department of Statistics, University of Glasgow, Glasgow G12 8QQ, UK*

Abstract

For two-class discrimination, Ref. [1] claimed that, when covariance matrices of the two classes were unequal, a (class) unbalanced dataset had a negative effect on the performance of linear discriminant analysis (LDA). Through re-balancing 10 real-world datasets, Ref. [1] provided empirical evidence to support the claim using AUC (Area Under the receiver operating characteristic Curve) as the performance metric. We suggest that such a claim is vague if not misleading, there is no solid theoretical analysis presented in [1], and AUC can lead to a quite different conclusion from that led to by misclassification error rate (ER) on the discrimination performance of LDA for unbalanced datasets. Our empirical and simulation studies suggest that, for LDA, the increase of the median of AUC (and thus the improvement of performance of LDA) from re-balancing is relatively small, while, in contrast, the increase of the median of ER (and thus the decline in performance of LDA) from re-balancing is relatively large. Therefore, from our study, there is no reliable empirical evidence to support the claim that a (class) unbalanced data set has a negative effect on the performance of LDA. In addition, re-balancing affects the performance of LDA for datasets with either equal or unequal covariance matrices, indicating that having unequal covariance matrices is not a key reason for the difference in performance between original and re-balanced data.

Preprint submitted to Pattern Recognition *22nd June 2007*

Key words: Area under an ROC curve (AUC); Linear discriminant analysis (LDA); Misclassification error rate (ER); Unbalanced data

1 Introduction

For two-class discrimination, Ref. [1] claims that, when covariance matrices of the two classes are unequal, a (class) unbalanced data set has a negative effect on the performance of linear discriminant analysis (LDA). We suggest that such a claim is vague if not misleading and we could find no solid theoretical analysis presented in [1]. However, their results of empirical experiments are interesting in finding that the performance of LDA on balanced data sets are superior to those of LDA on unbalanced data sets.

Following the notation used by [1], there are $n = n_1 + n_2$ observations with d features in the training set, where $\{\mathbf{x}_{1i}\}_{i=1}^{n_1}$ arise from class ω_1 and $\{\mathbf{x}_{2i}\}_{i=1}^{n_2}$ arise from class ω_2.

The Gaussian-based discrimination assumes two normal distributions: $(\mathbf{x}|\omega_1) \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $(\mathbf{x}|\omega_2) \sim \mathcal{N}(\mu_2, \Sigma_2)$ such that, for $j = 1, 2$,

$$g_j(\mathbf{x}) = \log(p(\mathbf{x}, \omega_j)) = -\frac{1}{2}(\mathbf{x}-\mu_j)^T\Sigma_j^{-1}(\mathbf{x}-\mu_j) - \frac{1}{2}\log|\Sigma_j| - \frac{d}{2}\log 2\pi + \log p(\omega_j) \,,$$

where $p(\omega_j)$ is the prior probability of class ω_j; it is a quadratic function of $\mathbf{x}$. When we assume further a common covariance matrix such that $\Sigma_1 = \Sigma_2 = \Sigma$, although $g_j(\mathbf{x})$ is still quadratic in $\mathbf{x}$ (not linear as stated in [1]), a discriminant

* Corresponding author. Tel.: +44 141 330 2474; fax: +44 141 330 4814.
Email addresses: jinghao@stats.gla.ac.uk (Jing-Hao Xue),
mike@stats.gla.ac.uk (D. Michael Titterington).

function $g^L(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$ becomes linear in $\mathbf{x}$. Consequently, Gaussian-based LDA is derived: $g^L(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$, where $\mathbf{w} = \Sigma^{-1}(\mu_1 - \mu_2)$, and

$$w_0 = \log\frac{p(\omega_1)}{p(\omega_2)} - \frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2) = \log\frac{p(\omega_1)}{p(\omega_2)} - \frac{1}{2}(\mu_1+\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2)\ .$$

Therefore, the optimal or Bayes discriminant rule of Gaussian-based LDA is to classify $\mathbf{x}$ into ω_1 if $\mathbf{w}^T\mathbf{x} + w_0 \geq 0$, and into ω_2 otherwise.

In practice, plug-in sample Gaussian-based LDA is commonly adopted by using relative frequencies of samples $\hat{p}(\omega_j) = n_j/(n_1 + n_2)$ to estimate $p(\omega_j)$, using sample means $\hat{\mu}_j$ to estimate μ_j, using sample within-class covariance matrices S_j to estimate Σ_j and using the pooled sample covariance matrix S to estimate Σ, where

$$\begin{aligned} S &= \frac{1}{n-2}\left(\sum_{i=1}^{n_1}(\mathbf{x}_{1i} - \hat{\mu}_1)(\mathbf{x}_{1i} - \hat{\mu}_1)^T + \sum_{i=1}^{n_2}(\mathbf{x}_{2i} - \hat{\mu}_2)(\mathbf{x}_{2i} - \hat{\mu}_2)^T\right) \\ &= \frac{1}{n-2}\left\{(n_1 - 1)S_1 + (n_2 - 1)S_2\right\}\ . \end{aligned}$$

Fisher's linear discriminant rule is to classify $\mathbf{x}$ into ω_1 if $\mathbf{w}^T\mathbf{x} \geq c$, where $\mathbf{w}^T\mathbf{x}$ is a linear combination of $\mathbf{x}$ and the coefficients $\mathbf{w}^T$ maximise the ratio $(\mathbf{w}^T\hat{\mu}_1 - \mathbf{w}^T\hat{\mu}_2)^2/(\mathbf{w}^T S\mathbf{w})$; the ratio is of the separation of the sample means of $\mathbf{w}^T\mathbf{x}$ to the pooled sample variance of $\mathbf{w}^T\mathbf{x}$. Differentiation of this ratio with respect to $\mathbf{w}$ results in $\mathbf{w} = \alpha S^{-1}(\hat{\mu}_1 - \hat{\mu}_2)$, where α is a scalar related to $n_j, \hat{\mu}_j$ and S (not necessarily $n-2$ alone as in [1]). Traditionally α is set to be 1 with the threshold c being adapted accordingly.

Fisher's linear discriminant rule does not assume Gaussian distributions for $\mathbf{x}|\omega_1$ and $\mathbf{x}|\omega_2$. However, in theory, it is equivalent to plug-in sample Gaussian-based LDA if the data satisfy the assumptions underlying the latter; in practice, it can be equivalent to the latter with $c = -w_0$. However, when the

assumptions underlying Gaussian-based LDA do not hold, for instance if $\Sigma_1 \neq \Sigma_2$, the optimal threshold c for a minimum classification error rate is not equal to $-w_0$ [2], and hence Fisher's linear discriminant rule differs from Gaussian-based LDA.

With the above formulae for Gaussian-based LDA, Ref. [1] claims that "if the two sample covariance matrices are different, the huge imbalance in class distribution is very problematic for LDA because the prior probability of majority class overshadows the differences in the sample covariance matrix terms. That is, the imbalanced data sets may hinder the performance of LDA". Such a claim is supported by their experimental results using re-balanced data obtained from original unbalanced data from four sampling methods [1].

2 Comments on the Claim

We suggest that the above mentioned claim and the empirical study to support it are vague if not misleading, even under an "ideal" condition such that $\hat{\mu}_j$ and S_j perfectly estimate μ_j and Σ_j, respectively. Let us explain it on three aspects.

First, if the true prior probabilities are approximately balanced such that $p(\omega_1) \approx p(\omega_2) \approx 0.5$ but the training set is unbalanced such that $n_1 \gg n_2$, then plug-in estimates $\hat{p}(\omega_j)$ are poor estimates of $p(\omega_j)$ because $\hat{p}(\omega_1) \gg \hat{p}(\omega_2)$, even though when the two sample covariance matrices are identical S will be a good estimate of Σ. Consequently, because of being based on $\frac{\hat{p}(\omega_1)}{\hat{p}(\omega_2)}$, w_0 is wrongly estimated so that LDA performs poorly. In this case, the use of re-balanced data, as in [1], will no doubt adjust $\hat{p}(\omega_j)$ such that $\hat{p}(\omega_j) \approx 0.5$ and

thus improve the performance of LDA. However, in practice, the training set is always given while the true priori probabilities are neither known nor necessarily balanced, and therefore the preprocessing of re-balancing data cannot guarantee a better performance of LDA.

Second, if the true prior probabilities are unbalanced such that $p(\omega_1) \gg p(\omega_2)$ and the training set demonstrates the imbalance such that $n_1 \gg n_2$, then plug-in estimates $\hat{p}(\omega_j) \approx p(\omega_j)$ are good estimates of $p(\omega_j)$ and thus $S = \hat{p}(\omega_1)S_1 + \hat{p}(\omega_2)S_2$ approaches the pooled population (within-class) covariance matrix $\Sigma = p(\omega_1)\Sigma_1 + p(\omega_2)\Sigma_2$. When the two sample covariance matrices are different, such that $S_1 \neq S_2$, the weights $\hat{p}(\omega_j)$ truly reflect the contribution of Σ_j to Σ. In contrast, if the training set is re-balanced by sampling as in [1], then $\hat{p}(\omega_j) = \frac{1}{2}$ are poor estimates of $p(\omega_j)$ and $S = \frac{1}{2}(S_1 + S_2)$. There is no reason to suggest that an LDA that uses $S = \frac{1}{2}(S_1 + S_2)$ and a wrongly estimated w_0 (with the term $\log \frac{\hat{p}(\omega_1)}{\hat{p}(\omega_2)} = 0$) will perform better than LDA that uses $S = \hat{p}(\omega_1)S_1 + \hat{p}(\omega_2)S_2$ where $\hat{p}(\omega_j) \approx p(\omega_j)$. Even if we assume that Ref. [1] uses accurate estimates of the prior probabilities $\hat{p}(\omega_j)$ from the original data such that $\hat{p}(\omega_j) \approx p(\omega_j)$ and uses the re-balanced data to estimate the pooled covariance matrix such that $S = \frac{1}{2}(S_1 + S_2)$ for Gaussian-based LDA, there is still no justification that such a linear classifier will approach the performance of the best "admissible" linear procedure under the condition that $\Sigma_1 \neq \Sigma_2$ [3], which is similar to Fisher's linear discriminant but with $\mathbf{w} = (t_1\Sigma_1 + t_2\Sigma_2)^{-1}(\mu_1 - \mu_2)$ (or in practice using sample statistics such that $\mathbf{w} = (t_1S_1 + t_2S_2)^{-1}(\hat{\mu}_1 - \hat{\mu}_2)$), where desired values of the scalars t_1 and t_2 have no closed-form solution so that systematic trials or computing algorithms have to be adopted [3,4,5].

Third, the misclassification error rate (ER) can be written as

$$\text{ER} = p(\omega_1)P(\omega_2|\omega_1) + p(\omega_2)P(\omega_1|\omega_2) \ ,$$

where $P(\omega_j|\omega_k)$ is the probability of misclassifying an observation, who arises from class k, into class j. For plug-in sample Gaussian-based LDA, when $(\mathbf{x}|\omega_1) \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $(\mathbf{x}|\omega_2) \sim \mathcal{N}(\mu_2, \Sigma_2)$, it follows that,

$$P(\omega_2|\omega_1) = P\left(\mathbf{w}^T\mathbf{x} + w_0 < 0 | \mathbf{x} \sim \mathcal{N}(\mu_1, \Sigma_1)\right) \ ,$$

$$P(\omega_1|\omega_2) = P\left(\mathbf{w}^T\mathbf{x} + w_0 \geq 0 | \mathbf{x} \sim \mathcal{N}(\mu_2, \Sigma_2)\right) \ .$$

Similarly to [4], the estimated probabilities of misclassification can be rewritten as

$$P(\omega_2|\omega_1) = \Phi\left(\frac{-\log\frac{\hat{p}(\omega_1)}{\hat{p}(\omega_2)} - \frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_2)^T S^{-1}(\hat{\mu}_1 - \hat{\mu}_2)}{[(\hat{\mu}_1 - \hat{\mu}_2)^T S^{-1}\Sigma_1 S^{-1}(\hat{\mu}_1 - \hat{\mu}_2)]^{\frac{1}{2}}}\right) = \Phi\left(-\frac{\mathbf{w}^T\hat{\mu}_1 + w_0}{\sqrt{\mathbf{w}^T\Sigma_1\mathbf{w}}}\right) \ ,$$

$$P(\omega_1|\omega_2) = \Phi\left(\frac{\log\frac{\hat{p}(\omega_1)}{\hat{p}(\omega_2)} - \frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_2)^T S^{-1}(\hat{\mu}_1 - \hat{\mu}_2)}{[(\hat{\mu}_1 - \hat{\mu}_2)^T S^{-1}\Sigma_2 S^{-1}(\hat{\mu}_1 - \hat{\mu}_2)]^{\frac{1}{2}}}\right) = \Phi\left(\frac{\mathbf{w}^T\hat{\mu}_2 + w_0}{\sqrt{\mathbf{w}^T\Sigma_2\mathbf{w}}}\right) \ ,$$

where Φ is the cumulative distribution function (CDF) of the standard normal distribution $\mathcal{N}(0, 1)$. Therefore, in the formula for ER, $p(\omega_j)$ and Σ_j are population parameters, or sample parameters from a sufficiently large original dataset, while $\hat{p}(\omega_j)$, $\hat{\mu}_j$ and S are sample statistics obtained from a training set.

In the experiments performed by [1], the test set includes $\frac{n_1}{4}$ observations arising from ω_1 and $\frac{n_2}{4}$ from ω_2 such that it conforms to the original relative frequencies; the remaining 75% of observations are then re-sampled into a training set with approximately equal number of observations from each class. Without explicit indication in [1] of how they obtain the sample relative frequencies $\hat{p}(\omega_j)$ (from the re-balanced training set or from the original data set) and the weights in calculating the pooled sample covariance matrix in those

experiments, we assume that all the parameters of the linear discriminant function are estimated from the re-balanced training set such that $\hat{p}(\omega_j) \approx \frac{1}{2}$ and $S \approx \hat{p}(\omega_1)S_1 + \hat{p}(\omega_2)S_2 = \frac{1}{2}(S_1 + S_2)$. In this context, a claim that using the re-balanced data can reduce ER can be translated into the following equality:

$$\frac{1}{2} = \underset{\hat{p}(\omega_1)}{\operatorname{argmin}} \left\{ p(\omega_1) P(\omega_2|\omega_1; \hat{p}(\omega_1)) + p(\omega_2) P(\omega_1|\omega_2; \hat{p}(\omega_1)) \right\} \ .$$

In order to verify this equality, we first perform some numerical evaluations on two specific scenarios: one is with $\Sigma_1 = \Sigma_2$, the other is with $\Sigma_1 \neq \Sigma_2$. In each scenario, we assume the original dataset is unbalanced with $p(\omega_1) = 0.8$, and there are large number of observations in both the test set and the training set such that $\hat{\mu}_j$ and S_j perfectly estimate μ_j and Σ_j, respectively, whether the data in the training set are unbalanced or balanced. With the population parameters $p(\omega_j)$, μ_j and Σ_j known, ER becomes a function of $\hat{p}(\omega_1)$ alone:

$$\mathrm{ER}(\hat{p}(\omega_1)) = p(\omega_1) P(\omega_2|\omega_1; \hat{p}(\omega_1)) + p(\omega_2) P(\omega_1|\omega_2; \hat{p}(\omega_1)) \ ,$$

where

$$P(\omega_2|\omega_1; \hat{p}(\omega_1)) = \Phi \left(\frac{-\log \frac{\hat{p}(\omega_1)}{1-\hat{p}(\omega_1)} - \frac{1}{2}(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)}{[(\mu_1 - \mu_2)^T \Sigma^{-1} \Sigma_1 \Sigma^{-1} (\mu_1 - \mu_2)]^{\frac{1}{2}}} \right) \ ,$$

$$P(\omega_1|\omega_2; \hat{p}(\omega_1)) = \Phi \left(\frac{\log \frac{\hat{p}(\omega_1)}{1-\hat{p}(\omega_1)} - \frac{1}{2}(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)}{[(\mu_1 - \mu_2)^T \Sigma^{-1} \Sigma_2 \Sigma^{-1} (\mu_1 - \mu_2)]^{\frac{1}{2}}} \right) \ ,$$

in which $\Sigma = \hat{p}(\omega_1)\Sigma_1 + (1 - \hat{p}(\omega_1))\Sigma_2$.

Here we consider a simple case in which each observation only has one feature (*i.e.*, $d = 1$). The population parameters are known to be $p(\omega_1) = 0.8$, $\mu_1 = 1$, $\mu_2 = -1$, $\Sigma_1 = 1$ and $\Sigma_2 \in [0.2, 5.0]$. The relationship between $\mathrm{ER}(\hat{p}(\omega_1))$ and $\hat{p}(\omega_1)$ is drawn in the 3-Dimensional plot as a function of $\hat{p}(\omega_1)$ and Σ_2 in the left panel of Figure 1. The surface of $\mathrm{ER}(\hat{p}(\omega_1))$ does not have a minimum

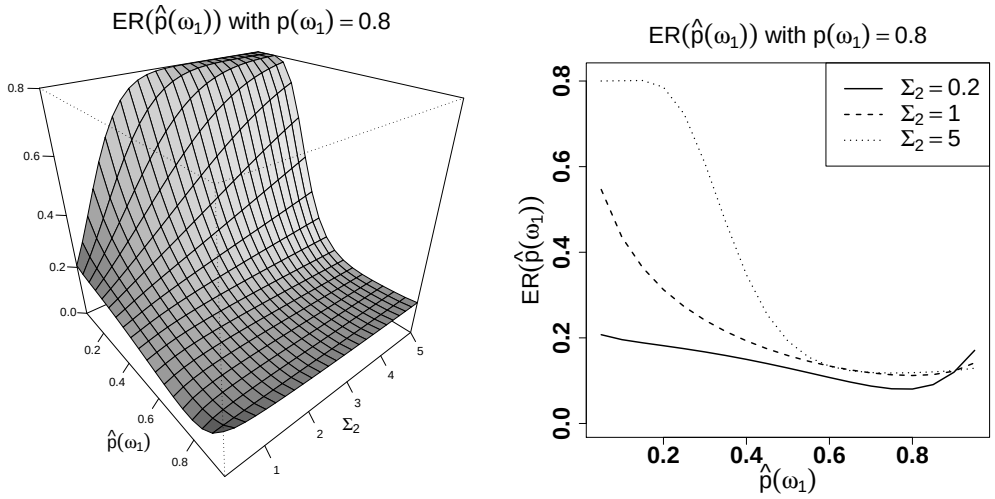

Figure 1. The misclassification error rates ER($\hat{p}(\omega_1)$).

point at $\hat{p}(\omega_1) = 0.5$.

In the right panel of Figure 1, we draw the curves of ER($\hat{p}(\omega_1)$) for $\Sigma_2 = 0.2, 1,$ and 5, respectively. We observe the following.

(1) When $\Sigma_2 = 0.2$ or 5 such that $\Sigma_2 \neq \Sigma_1$, the best performance of LDA is obtained at $\hat{p}(\omega_1) = 0.8$, which is equal to the true prior probability of class ω_1, rather than from the re-balanced data, which gives $\hat{p}(\omega_1) = 0.5$; the procedure of re-balancing data has a negative effect on the performance of LDA if the original unbalanced data conform to the truly unbalanced population.

(2) When $\Sigma_2 = 1$ such that $\Sigma_2 = \Sigma_1$, the best performance of LDA is also obtained at $\hat{p}(\omega_1) = 0.8$ rather than from the re-balanced data; the procedure of re-balancing data may also have a negative effect.

(3) In general, the data with a compact within-class distribution (in the sense of a small within-class covariance matrix) may result in a better performance of LDA (in the sense of smaller ER($\hat{p}(\omega_1)$)), compared with the data with a dispersed within-class distribution.

(4) In fact, in this case, since $\min(p(\omega_1), p(\omega_2)) = 0.2$, in practice the maximum $\mathrm{ER}(\hat{p}(\omega_1))$ can be controlled to be 0.2, the smaller prior probability, if we always classify observations into the class with higher prior probability.

In summary, under the condition of large number of observations, with regard to ER as the measure of performance, there is no evidence from our numerical evaluations to justify the claim that re-balancing original data can improve the performance of Gaussian-based LDA, and the best performance of LDA is always obtained when the estimated priori probabilities conforms to the true population prior probabilities.

3 AUC or ER

Unbalanced datasets are quite common in practice. For two-class discrimination, conventionally one of two classes which has higher prior probability is called the majority or negative class, and the other class is called the minority or positive class. In practice, many discrimination techniques are not very successful in identifying the minority class [6].

There are many approaches to dealing with data imbalance (rarity) [7]. The simplest approaches are random over-sampling with replacement and under-sampling, where the former is to increase the number of the minority class and the latter is to reduce the number of the majority class. Such sampling will modify the class distributions of the training data. Random over-sampling cannot gain new information about the minority class; random under-sampling may lose useful information about the majority class. Nevertheless, for practi-

cal datasets, such sampling may improve the performance of LDA with regard to certain evaluation metrics, as shown by [1].

The ER, also called "accuracy" in [8,9,7], is the most widely used evaluation metric for classifiers such as LDA. However, as an average over all the observations that are classified, it inevitably favours the majority class given the assumption that the error in the minority class is of equal importance to that in the majority class. Therefore, it can be biased by the prior probabilities if errors have in practice different importance between the two classes; it is recommended to use a loss function in this case.

For two-class discrimination of unbalanced data, where the error in the minority class may be more important in practice, the Receiver Operating Characteristic (ROC) curve and the area under the curve, the so-called AUC, are commonly used [9,7]. The ROC curve is a plot of the true positive rate vs. the false positive rate, and hence a higher AUC generally indicates a better classifier. As pointed out by [10], there is a three-way equivalence between AUC, the Wilcoxon-Mann-Whitney statistic and the probability of a correct ranking of a randomly chosen (negative, positive) pair. More precisely, suppose that a discriminant function such as $g^L(\mathbf{x})$ is designed to provide a high score for a positive observation and a low score for a negative one, then, given a randomly chosen (negative, positive) pair denoted by $(\mathbf{x}_N, \mathbf{x}_P)$, it holds that $\text{AUC} = Prob(\mathbf{x}_N < \mathbf{x}_P)$.

Such equivalence to the Wilcoxon-Mann-Whitney statistic is also mentioned in [8,11,1], and hence AUC is concerned more about ranking than about the misclassification error of the predictions [11]. In contrast to ER, AUC is invariant to the prior probabilities [8].

The ROC is obtained by varying the discriminant threshold, while, in practice, ER is obtained for some classifiers such as LDA at a conventionally fixed, discriminant threshold which is optimal under certain assumptions. Therefore, AUC is independent of the discriminant threshold while ER is not.

Concerning the relationship between AUC and ER, Ref. [8] shows that there is good agreement between these two evaluation metrics in ranking 9 classification algorithms including C4.5 (an algorithm based on classification trees) and plug-in sample Gaussian-based quadratic discriminant analysis (QDA). Furthermore, the theoretical analysis in [11] shows that the mean of AUC is monotonically decreasing as ER increases. Meanwhile, Ref. [11] shows that, the more unbalanced the data, the higher the coefficient of variation of AUC and the lower the mean of AUC. This not only indicates that AUC may suggest a different conclusion from that drawn by ER with regard to classifier performance on unbalanced data, but also suggests that using AUC as the evaluation metric favours balanced data. In fact, using C4.5, Ref. [12] presents a thorough empirical study of 26 real-world datasets; their results show that, in general, ER is better with original data while AUC is better with re-balanced data.

Ref. [1] uses AUC to evaluate the performance of plug-in sample Gaussian LDA (denoted by LDA-Σ hereafter); in our study, we will use both AUC and ER to evaluate the performance of LDA-Σ and one of its special versions which assumes that the common covariance matrix is diagonal (denoted by LDA-Λ). We first replicate the experiments in [1] on 10 datasets from the UCI machine learning repository [13] with our implementations, and then investigate 4 simulated datasets of normally distributed data and normal mixture data.

4 Replication of Experiments on UCI Datasets

As with [12] and [1], the test set is constructed by including $\frac{n_1}{4}$ observations arising from the minority class ω_1 and $\frac{n_2}{4}$ from the majority class ω_2 such that it maintains the prevalence rate of each class; the remaining 75% of observations in the original, unbalanced training set are then re-sampled into two training sets with equal numbers of observations from each class, respectively by random over-sampling with replacement and random under-sampling.

We implement such constructions randomly T times; such a validation is not a cross-validation since the training set and test set are not necessarily crossed over. However, it can be expected that such a validation is as effective as T-fold cross-validation, if T is a large number. In our implementation, $T = 200$. As suggested in [8], we average over the T AUCs to obtain one average AUC, rather than average over the T ROCs to calculate one AUC.

The AUC is obtained through calculating the Wilcoxon-Mann-Whitney statistic of the predicting scores for LDA. It is implemented by an R function *wilcox.test* from a standard package **stats** in R to perform the Mann-Whitney test (equivalently the Wilcoxon rank sum test) for two unpaired samples. In order to exercise the test, scores of the discriminant function $g^L(\mathbf{x})$ are used as the varying discriminant threshold and for ranking.

Table 1 presents the description of the 10 UCI datasets being studied (the class prior probabilities different from Table 1 of [1] are highlighted in italics).

As with [8] and [14], the UCI data are rescaled into the range $[0, 1]$. In addition, before carrying out LDA, we perform for each feature $\mathbf{x}_i|\mathbf{y}$ the Shapiro-Wilk

Data set	Observations	Features	Class (min., maj.)	Prior (min., maj.)
Letter-a	20,000	16	(A, remainder)	(3.94%, 96.06%)
Satimage-3	6,435	36	(3, remainder)	(21.1%, 78.9%)
Waveform	5,000	21	(1, remainder)	(*32.94%, 67.06%*)
Image	2,310	18	(BRICKFACE, remainder)	(14.29%, 85.71%)
Vehicle	846	18	(van, remainder)	(23.52%, 76.48%)
Pima	768	8	(1, 0)	(*34.9%, 65.1%*)
New-thyroid	215	5	(hypo, remainder)	(*13.95%, 86.05%*)
Glass	214	9	(3, remainder)	(7.94%, 92.06%)
Wine	178	13	(3, remainder)	(26.97%, 73.03%)
Iris	150	4	(Iris-virginica, remainder)	(33.33%, 66.67%)

Table 1

Description of data

test for within-class normality and Levene's test for homogeneity of variance across the two classes at the significance level 0.05. If for a feature the within-class normality is rejected in any of the two classes, we mark the feature as "Normality rejected". Results of these two tests, as shown in Table 2, suggest that for all 10 datasets under study the null hypotheses of within-class normality and homoscedasticity across the classes are rejected, including the dataset "Pima" which is stated to have nearly equal sample covariance matrices in [1].

Data set	Features	Normality rejected	Homoscedasticity rejected
Letter-a	16	16	12
Satimage-3	36	36	36
Waveform	21	15	15
Image	18	18	18
Vehicle	18	18	14
Pima	8	8	5
New-thyroid	5	5	3
Glass	9	9	2
Wine	13	12	10
Iris	4	3	3

Table 2

Results of the Shapiro-Wilk test for within-class normality and Levene's test for homogeneity of variance across the two classes.

Table 3, 4, 5 and 6 list our results, obtained from LDA-Σ and LDA-Λ, of medians of AUC and ER for the original and re-balanced data, as well as p-values for the Wilcoxon signed rank test for the pairs of (original, over-sampling) and of (original, under-sampling). From the tables, we can observe the following.

(1) Concerning LDA-Σ, AUCs of re-balanced data are significantly (at the

Data set	Original	Over.	Under.	p-v. (Ori.-Over.)	p-v. (Ori.-Under.)
Letter-a	0.977	0.986	0.986	0	0
Satimage-3	0.987	0.988	0.987	0	0
Waveform	0.943	0.945	0.944	0	0
Image	0.994	0.995	0.995	0	0
Vehicle	0.989	0.993	0.991	0	0
Pima	0.835	0.840	0.834	0	0.801
New-thyroid	0.995	1	0.997	0	0.083
Glass	0.827	0.918	0.801	0	0.018
Wine	1	1	1	0.005	0.01
Iris	0.977	0.990	0.987	0	0

Table 3

Results from LDA-Σ: medians of AUC for the original and re-balanced data and p-values for the Wilcoxon signed rank test for pairs of (original, over-sampling) and of (original, under-sampling).

level 0.05) better than those of original data, except for the under-sampled data of "Pima", "New-thyroid" and "Glass". Although the increase of its median (and thus the improvement of classifier performance) from re-balancing is not relatively large in amount, in general, it can be said that, for the datasets being studied, AUC favours re-balanced data.

(2) Concerning LDA-Λ: of the 10 datasets, AUCs of re-balanced "Satimage-

Data set	Original	Over.	Under.	p-v. (Ori.-Over.)	p-v. (Ori.-Under.)
Letter-a	0.011	0.044	0.045	0	0
Satimage-3	0.051	0.076	0.077	0	0
Waveform	0.126	0.170	0.171	0	0
Image	0.019	0.033	0.036	0	0
Vehicle	0.047	0.047	0.052	0.827	0.002
Pima	0.224	0.234	0.240	0	0
New-thyroid	0.056	0.019	0.037	0	0
Glass	0.075	0.226	0.292	0	0
Wine	0	0.023	0.023	0	0
Iris	0.081	0.108	0.108	0	0

Table 4

Results from LDA-Σ: medians of ER for the original and re-balanced data and p-values for the Wilcoxon signed rank test for pairs of (original, over-sampling) and of (original, under-sampling).

3" and "Image" are significantly worse than those of the original data for both re-sampling methods, and AUC of re-balanced "Glass" is significantly worse than that of original data for the under-sampling. Meanwhile, no significant difference exists between AUCs of "Vehicle". This may be because of the different estimates of the covariance matrix between LDA-Σ and LDA-Λ; this indicates that the accuracy of estimation

Data set	Original	Over.	Under.	p-v. (Ori.-Over.)	p-v. (Ori.-Under.)
Letter-a	0.951	0.952	0.952	0	0
Satimage-3	0.982	0.981	0.981	0	0
Waveform	0.916	0.917	0.917	0	0
Image	0.873	0.864	0.865	0	0
Vehicle	0.783	0.782	0.783	0.204	0.06
Pima	0.818	0.822	0.820	0	0.023
New-thyroid	0.997	1	1	0	0.136
Glass	0.709	0.750	0.653	0	0
Wine	1	1	1	0	0.096
Iris	0.990	0.990	0.990	0	0

Table 5

Results from LDA-Λ: medians of AUC for the original and re-balanced data and p-values for the Wilcoxon signed rank test for pairs of (original, over-sampling) and of (original, under-sampling).

can play a more important role in AUC than the re-balancing does.

(3) In contrast to AUC, ER is significantly increased by re-balancing except for "New-thyroid" and "Vehicle". The increase of its median (and thus the decline of classifier performance) from re-balancing is relatively large. In general, it can be said that, for the datasets being studied, ER favours original data.

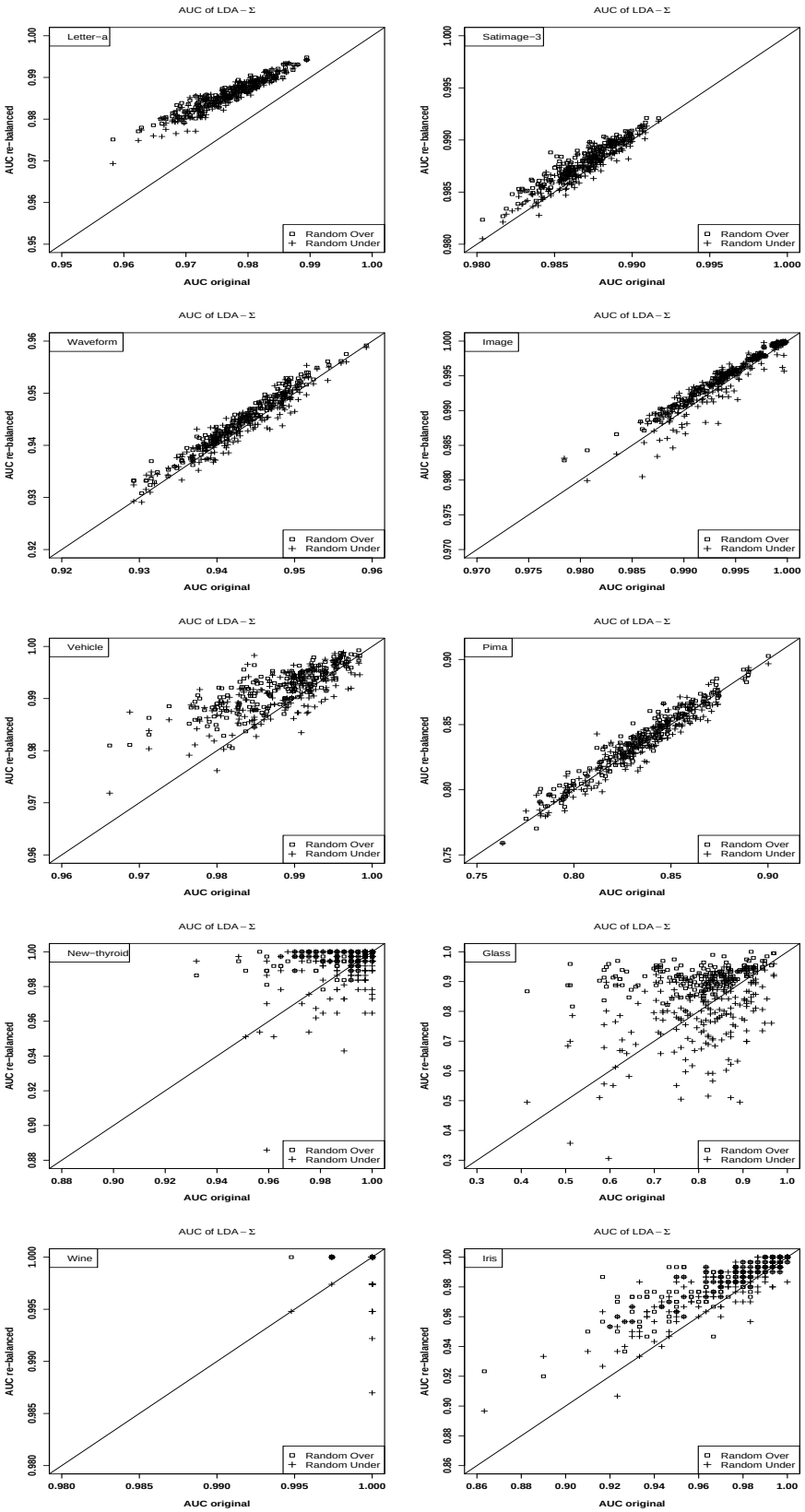

Figure 2. Scatter plots of AUC on re-balanced data (by over-sampling and under-sampling) vs. original data, obtained from LDA-Σ.

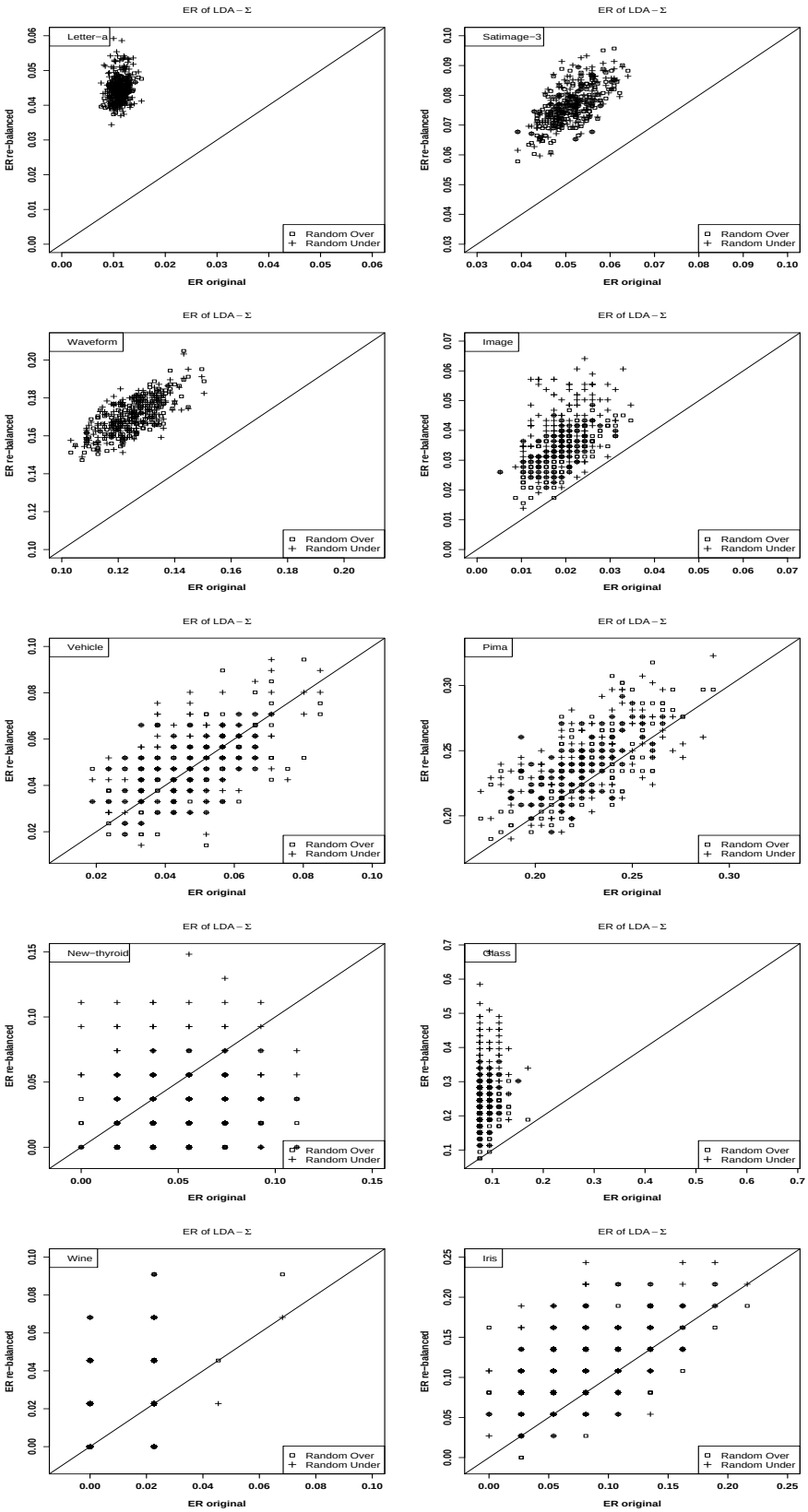

Figure 3. Scatter plots of ER on re-balanced data (by over-sampling and under-sampling) vs. original data, obtained from LDA-Σ.

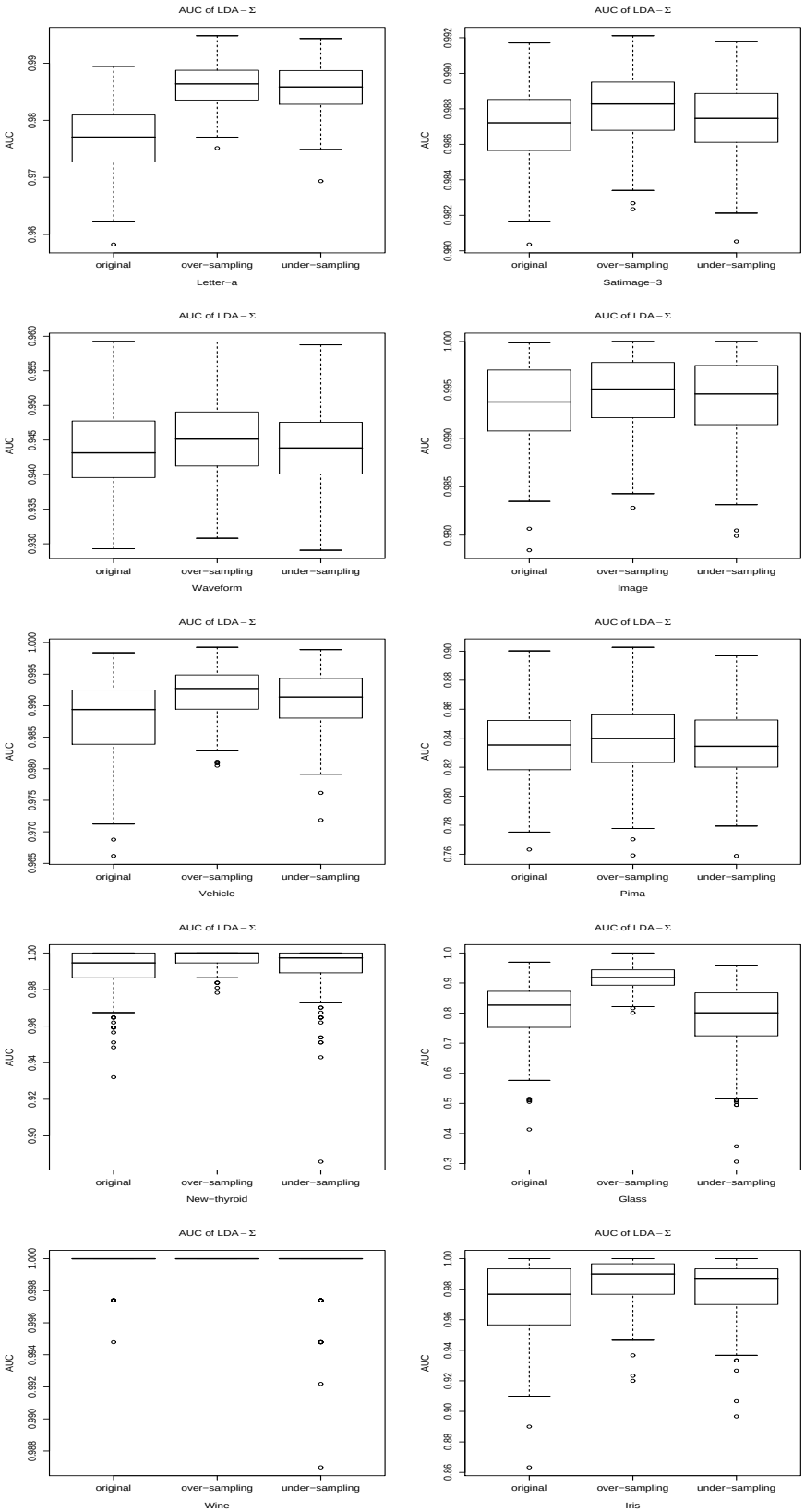

Figure 4. Box-plots of AUC on original and re-balanced data (by over-sampling and under-sampling), obtained from LDA-Σ.

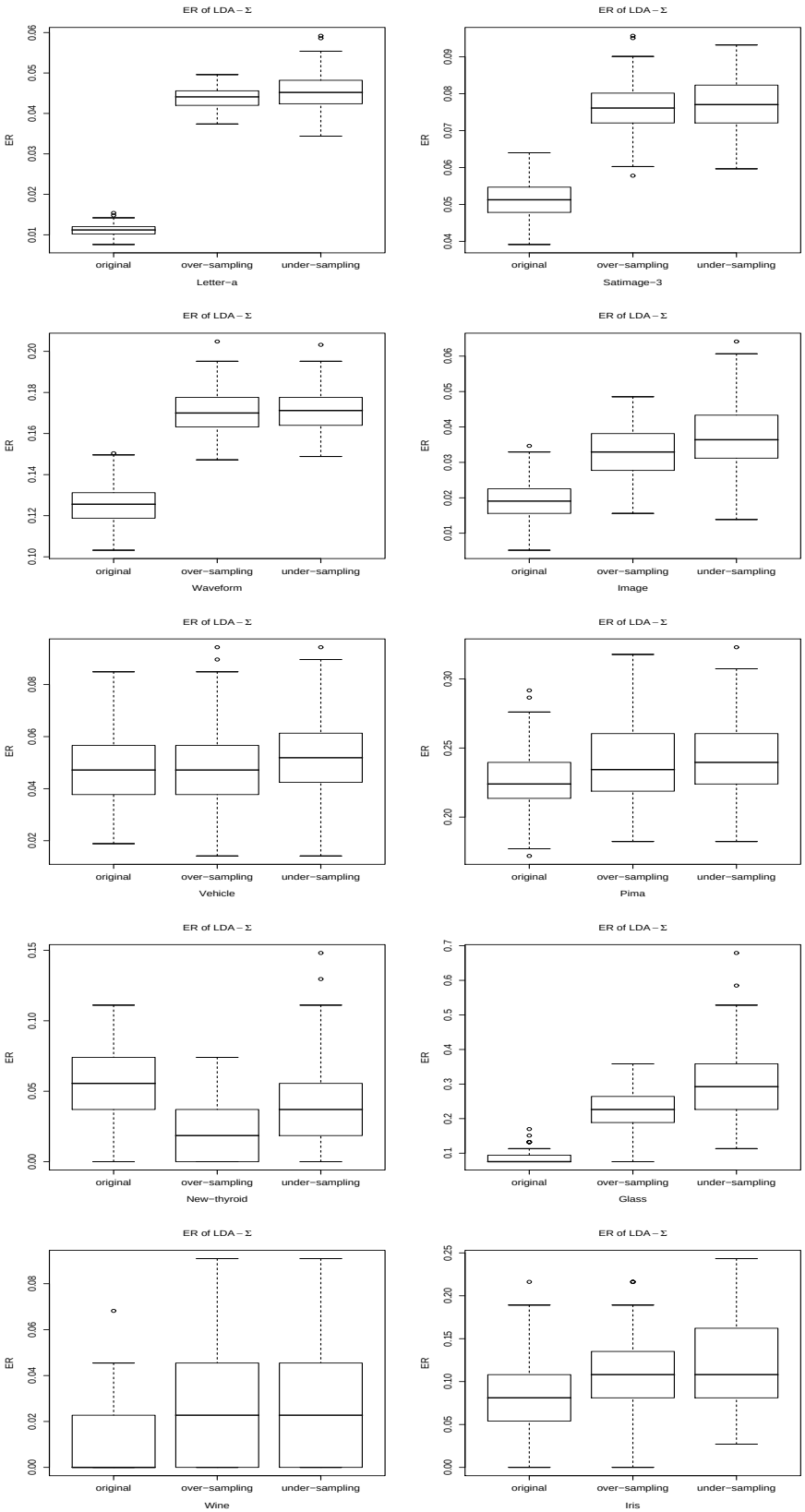

Figure 5. Box-plots of ER on original and re-balanced data (by over-sampling and under-sampling), obtained from LDA-Σ.

Data set	Original	Over.	Under.	p-v. (Ori.-Over.)	p-v. (Ori.-Under.)
Letter-a	0.023	0.076	0.076	0	0
Satimage-3	0.121	0.136	0.135	0	0
Waveform	0.154	0.163	0.163	0	0
Image	0.218	0.310	0.310	0	0
Vehicle	0.363	0.363	0.358	0.002	0.001
Pima	0.245	0.260	0.260	0	0
New-thyroid	0.037	0.019	0.019	0	0
Glass	0.075	0.509	0.509	0	0
Wine	0.023	0.045	0.045	0	0
Iris	0.135	0.162	0.162	0	0

Table 6

Results from LDA-Λ: medians of ER for the original and re-balanced data and p-values for the Wilcoxon signed rank test for pairs of (original, over-sampling) and of (original, under-sampling).

Obtained from LDA-Σ on the 10 datasets, scatter plots of AUC and ER on re-balanced (by over-sampling and under-sampling) vs. original data are shown in Figures 2 and 3, and box-plots of AUC and ER on original and re-balanced data are shown in Figures 4 and 5, respectively. Results from LDA-Λ are similar and thus are omitted here.

5 Simulation Studies

Although we may observe some patterns from the empirical study using real-world datasets such as those from the UCI machine learning repository, it is not reliable to generalise the patterns into a conclusion beyond the tested datasets. In this sense, a study on simulated datasets can be a good complement to the empirical study.

In [4], simulation studies by Monte Carlo methods are used to compare the performance of the so-called best linear function [3], the quadratic and Fisher's linear discriminant function, under the condition that $\Sigma_1 \neq \Sigma_2$. One of the simulation studies with respect to $p(\omega_j)$ and $\hat{p}(\omega_j)$ shows that ER is smaller when $\hat{p}(\omega_j)$ is closer to $p(\omega_j)$.

Fisher's linear discriminant rule as used in [4] is in fact a variant of the plug-in sample Gaussian-based LDA with $\mathbf{w} = S^{-1}(\hat{\mu}_1 - \hat{\mu}_2)$, and

$$w_0 = \log \frac{p(\omega_1)}{p(\omega_2)} - \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)^T S^{-1}(\hat{\mu}_1 - \hat{\mu}_2) \ ,$$

where population prior probabilities $p(\omega_j)$ are used for the term $\log \frac{p(\omega_1)}{p(\omega_2)}$ in w_0 while sample prior probabilities $\hat{p}(\omega_j)$ are used in $S = \hat{p}(\omega_1)S_1 + \hat{p}(\omega_2)S_2$. In practice, since the $p(\omega_j)$ are unknown, $\log \frac{\hat{p}(\omega_1)}{\hat{p}(\omega_2)}$ is more widely used in w_0.

In this section, we simulate 4 datasets; each dataset consists of 1000 observations and is divided into two classes, ω_1 and ω_2, with 200 observations from the minority class ω_1 and 800 observations from the majority class ω_2 such that each dataset is unbalanced with $\hat{p}(\omega_1) = 0.2$. The first dataset is randomly generated from two 4-variate normal distributions, $\mathbf{x}|\omega_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $\mathbf{x}|\omega_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$, with equal covariance matrices such that $\Sigma_1 = \Sigma_2$;

the second dataset is similar to the first one except that $\Sigma_1 \neq \Sigma_2$. The third and fourth datasets are randomly generated from two 4-variate normal mixtures; each mixture has two components. The third one has equal covariance matrices across the two classes while the fourth one does not.

For $\mathbf{x}|\omega_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $\mathbf{x}|\omega_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$, as with [4] and [2], we can use a linear transformation to reduce Σ_1 to the identity matrix $\mathbf{I}$ and diagonalise Σ_2. Therefore, without loss of generality, in this section, we use a canonical form with $\mu_1 = \mathbf{0}$, $\Sigma_1 = \mathbf{I}$ and $\mu_2 = (-1.5, -0.75, 0.75, 1.5)^T$, and with Σ_2 a diagonal covariance matrix. For the dataset with equal covariance matrices, $\Sigma_2 = \mathbf{I} = \Sigma_1$; for the dataset with unequal covariance matrices, Σ_2 is a diagonal matrix with 4 diagonal elements which are $(0.25, 0.75, 1.25, 1.75)$, so that $\Sigma_2 \neq \Sigma_1$.

Compared with the normal distribution, the mixture of normal distributions is a better approximation to real data in a variety of situations. In this section, 2 simulated datasets are randomly generated from two mixtures, ω_1 and ω_2, of 4-variate normal distributions.

Each mixture has two components with equal mixing coefficients. The two components, A and B, of the mixture ω_1 are normally distributed with probability density functions $\mathcal{N}(\mu_{1A}, \Sigma_1)$ and $\mathcal{N}(\mu_{1B}, \Sigma_1)$, respectively, where $\mu_{1A} = \mathbf{0}$ and $\mu_{1B} = (2, 0, 0, 0)^T$; and the two components, C and D, of the mixture ω_2 are normally distributed with probability density functions $\mathcal{N}(\mu_{2C}, \Sigma_2)$ and $\mathcal{N}(\mu_{2D}, \Sigma_2)$, respectively, where $\mu_{2C} = (-1.5, -0.75, 0.75, 1.5)^T$ and $\mu_{2D} = (-3.5, -0.75, 0.75, 1.5)^T$. In such a way, when Σ_1 and Σ_2 are equal/unequal, the covariance matrices of the two mixtures will become equal/unequal. Meanwhile, we set Σ_1 and Σ_2 in the same way as for the normally distributed data.

In our simulation studies, both Σ_1 and Σ_2 are diagonal; the performance of

LDA-Λ is found similar to that of LDA-Σ, and thus only the results obtained from LDA-Σ are presented in the following.

The simulations from the multivariate normal distributions and normal mixtures are based on an R function *mvrnorm* for simulating from a contributed R package **MASS**. As with the UCI datasets being studied, the simulated data are rescaled into the range $[0, 1]$.

Data set	Original	Over.	Under.	p-v. (Ori.-Over.)	p-v. (Ori.-Under.)
Normal-equ	0.962	0.963	0.962	0	0.012
Normal-unequ	0.943	0.949	0.948	0	0
Mixture-equ	0.981	0.982	0.981	0	0.26
Mixture-unequ	0.992	0.992	0.992	0.151	0.001

Table 7

Results from LDA-Σ: medians of AUC for the original and re-balanced data and p-values for the Wilcoxon signed-rank test for pairs of (original, over-sampling) and of (original, under-sampling).

Table 7 and 8 list our results, obtained from LDA-Σ, of medians of AUC and ER for the original and re-balanced data, as well as p-values for the Wilcoxon signed-rank test for the pairs of (original, over-sampling) and of (original, under-sampling). From the tables, we can observe the following.

(1) Concerning AUC obtained from both LDA-Σ, although for the simulated datasets being studied it generally favours re-balanced data, the increase of its median (and thus the improvement of performance of LDA) from

Data set	Original	Over.	Under.	p-v. (Ori.-Over.)	p-v. (Ori.-Under.)
Normal-equ	0.072	0.108	0.112	0	0
Normal-unequ	0.060	0.096	0.096	0	0
Mixture-equ	0.056	0.068	0.068	0	0
Mixture-unequ	0.032	0.044	0.044	0	0

Table 8

Results from LDA-Σ: medians of ER for the original and re-balanced data and p-values for the Wilcoxon signed-rank test for pairs of (original, over-sampling) and of (original, under-sampling).

re-balancing is relatively small. We observe that, of the two simulated mixture datasets, there is no significant change in AUC between under-sampled and original data for one dataset and between over-sampled and original data for the other dataset.

(2) Concerning ER obtained from both LDA-Σ, in contrast to AUC, all ERs are significantly increased after the data are re-balanced. ER favours original data and the increase of its median (and thus the decline in performance of LDA) from re-balancing is noticeably large.

Obtained from LDA-Σ on the 4 simulated datasets, scatter plots of AUC and ER on re-balanced (by over-sampling and under-sampling) vs. original data are shown in Figures 6 and 7, and box-plots of AUC and ER on original and re-balanced data are shown in Figures 8 and 9, respectively.

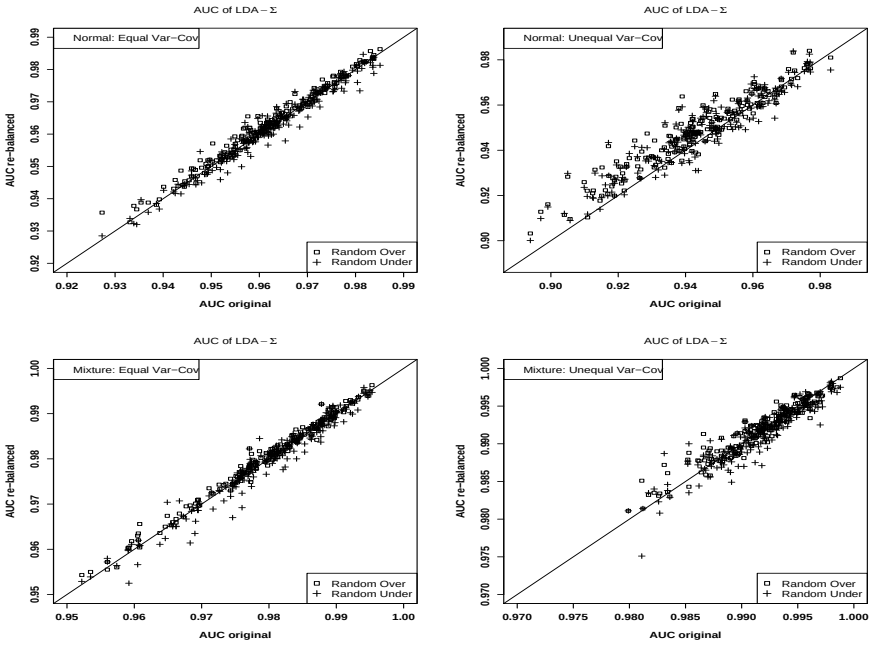

Figure 6. Scatter plots of AUC on re-balanced data (by over-sampling and under-sampling) vs. original data, obtained from LDA-Σ.

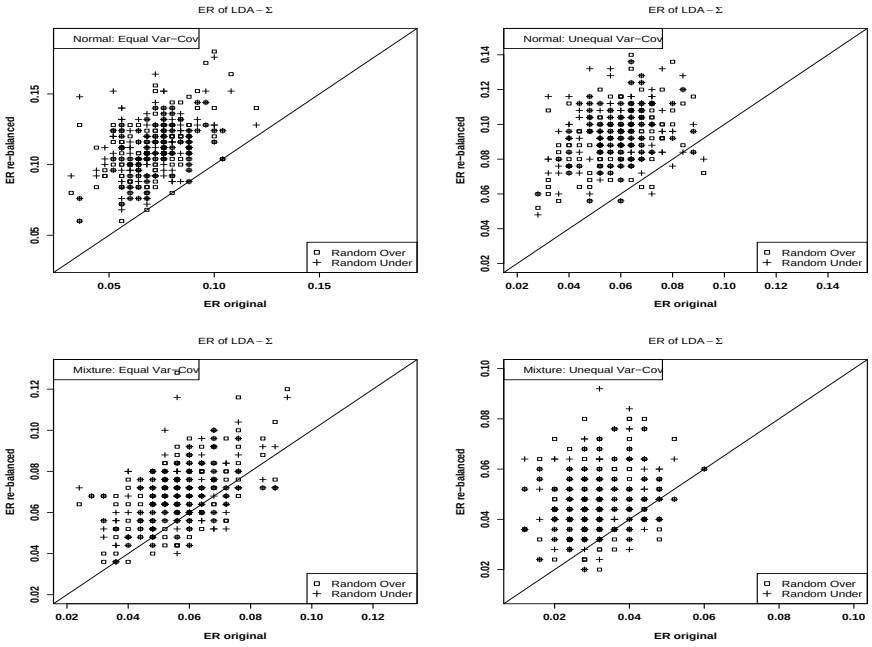

Figure 7. Scatter plots of ER on re-balanced data (by over-sampling and under-sampling) vs. original data, obtained from LDA-Σ.

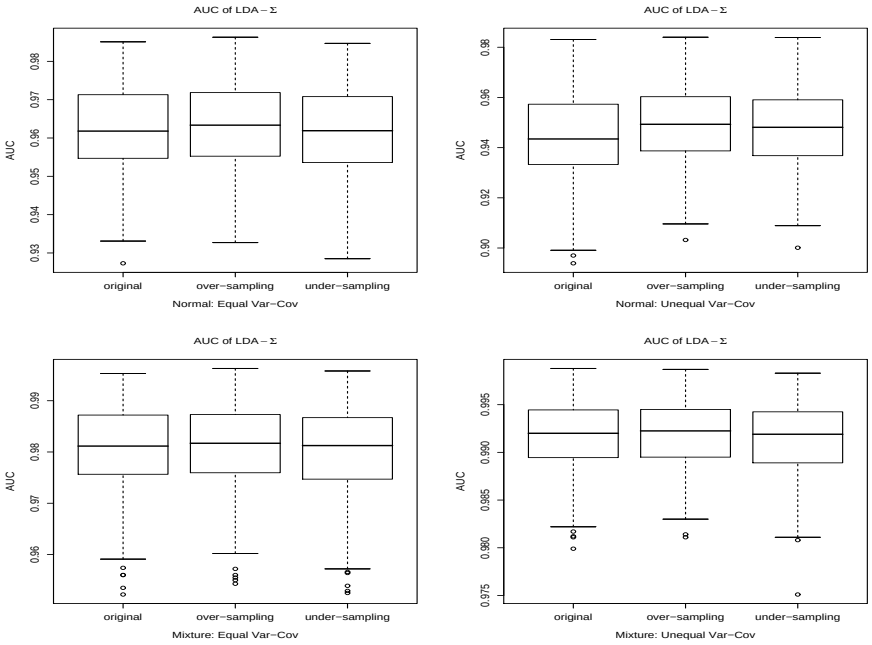

Figure 8. Box-plots of AUC on original and re-balanced data (by over-sampling and under-sampling), obtained from LDA-Σ.

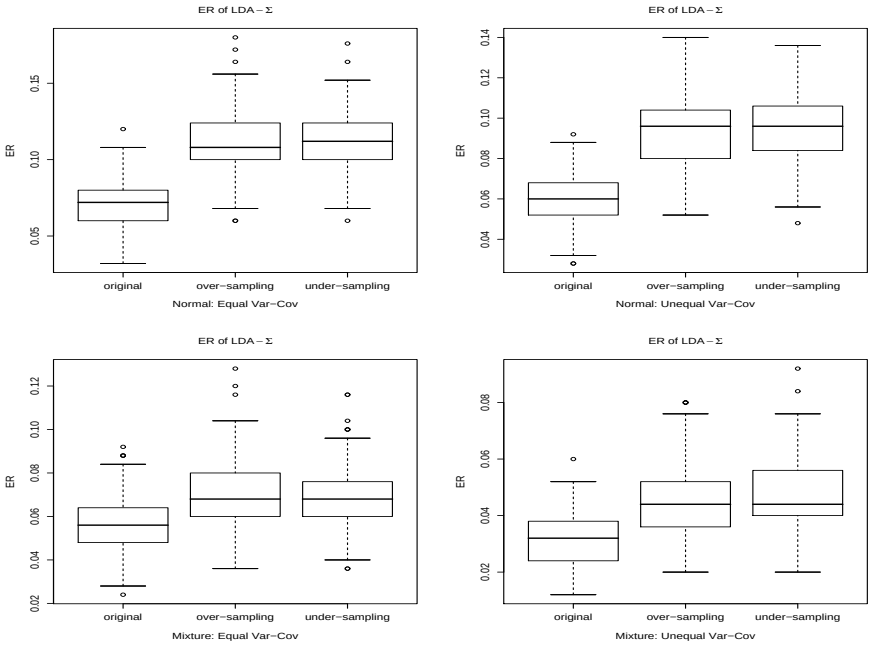

Figure 9. Box-plots of ER on original and re-balanced data (by over-sampling and under-sampling), obtained from LDA-Σ.

6 Conclusions

In general, we can draw the following conclusions with regard to the datasets in our study.

(1) Concerning AUC obtained from LDA, although it generally favours re-balanced data, the increase of its median (and thus the improvement of performance of LDA) from re-balancing is relatively small. In contrast to AUC, ER favours original data and the increase of its median (and thus the decline in performance of LDA) from re-balancing is relatively large. This shows that AUC and ER can lead to quite different conclusions on the discrimination performance of LDA for unbalanced datasets.

(2) Therefore, from our study, there is no reliable empirical evidence to support the claim that a (class) unbalanced data set has a negative effect on the performance of LDA.

(3) Re-balancing affects the performance of LDA for both the datasets with equal or unequal covariance matrices. This indicates that having unequal covariance matrices is not a key reason for the difference in performance between original and re-balanced data.

References

[1] J. G. Xie, Z. D. Qiu, The effect of imbalanced data sets on LDA: a theoretical and empirical analysis, Pattern Recognition 40 (2) (2007) 557–562.

[2] E. S. Gilbert, The effect of unequal variance-covariance matrices on Fisher's linear discriminant function, Biometrics 25 (3) (1969) 505–515.

[3] T. W. Anderson, R. R. Bahadur, Classification into two multivariate normal

distributions with different covariance matrices, The Annals of Mathematical Statistics 33 (2) (1962) 420–431.

[4] S. Marks, O. J. Dunn, Discriminant function when covariance matrices are unequal, Journal of the American Statistical Association 69 (345) (1974) 555–559.

[5] G. J. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, John Wiley & Sons, New York, 1992.

[6] D. M. Titterington, G. D. Murray, L. S. Murray, D. J. Spiegelhalter, A. M. Skene, J. D. F. Habbema, G. J. Gelpke, 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) (1981) 145–175.

[7] G. M. Weiss, Mining with rarity: a unifying framework, SIGKDD Explorations 6 (1) (2004) 7–19.

[8] A. P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition 30 (7) (1997) 1145–1159.

[9] N. V. Chawla, N. Japkowicz, A. Kotcz, Editorial: special issue on learning from imbalanced data sets, SIGKDD Explorations 6 (1) (2004) 1–6.

[10] J. A. Hanley, B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1) (1982) 29–36.

[11] C. Cortes, M. Mohri, AUC optimization vs. error rate minimization, in: NIPS, 2003.

[12] G. M. Weiss, F. Provost, Learning when training data are costly: the effect of class distribution on tree induction, Journal of Artificial Intelligence Research 19 (2003) 315–354.

[13] D. J. Newman, S. Hettich, C. L. Blake, C. J. Merz, UCI Repository of machine learning databases, University of California, Irvine, Dept. of Information and Computer Sciences, http://www.ics.uci.edu/∼mlearn/MLRepository.html (1998).

[14] A. Y. Ng, M. I. Jordan, On discriminative vs. generative classifiers: a comparison of logistic regression and naïve bayes, in: NIPS, 2001, pp. 841–848.

Jing-Hao Xue

Jing-Hao Xue was born in Jiangxi, China, in 1971. He received the B.Eng degree in Telecommunication and Information Systems in 1993 and the Dr.Eng degree in Signal and Information Processing in 1998, both from Tsinghua University, and the MSc degree in Medical Imaging and the MSc degree in Statistics, both from Katholieke Universiteit Leuven in 2004.

He is currently a Ph.D. student in Statistics in the Department of Statistics at the University of Glasgow. His research interests include statistical pattern recognition including generative and discriminative modelling.

D. Michael Titterington

D. Michael Titterington was born in Marple, England, in 1945. He received the B.Sc. degree in mathematical science from the University of Edinburgh in 1967 and the degree of Ph.D. from the University of Cambridge in 1972.

He has worked in the Department of Statistics at the University of Glasgow since 1972; he was appointed Titular Professor in 1982 and Professor in 1988. He was Head of Department from 1982 until 1991. He has held visiting appointments at Princeton University, SUNY at Albany, the University of Wisconsin-Madison and the Australian National University. His research interests Include optimal design, incomplete data problems including mixtures, statistical pattern recognition, statistical smoothing, including image analysis, and statistical aspects of neural networks.

Dr. Titterington was elected Fellow of the Institute of Mathematical Statistics in 1996, Member of the International Statistics Institute in 1991, and Fellow of the Royal Society of Edinburgh, also in 1991. He has held editorial appointment with the *Annals of Statistics, Biometrika, the Journal of the American Statistical Association*, the *Journal of the Royal Statistical Society (Series B), Statistical Science*, and *IEEE Transactions on Pattern Analysis and Machine Intelligence*.