

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.

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 from empirical experiments are interesting in finding that the performance of LDA on balanced data sets is superior to that of LDA on unbalanced data sets.

In 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 .

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 Titterton).

1 function $g^L(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$ becomes linear in \mathbf{x} . Consequently, Gaussian-
 2 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

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

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

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

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

$$= \frac{1}{n-2} \{ (n_1 - 1)S_1 + (n_2 - 1)S_2 \} .$$

11 Fisher's linear discriminant rule is to classify \mathbf{x} into ω_1 if $\mathbf{w}^T \mathbf{x} \geq c$, where
 12 $\mathbf{w}^T \mathbf{x}$ is a linear combination of \mathbf{x} and the coefficients \mathbf{w}^T maximise the ratio
 13 $(\mathbf{w}^T \hat{\mu}_1 - \mathbf{w}^T \hat{\mu}_2)^2 / (\mathbf{w}^T S \mathbf{w})$; the ratio is of the separation of the sample means
 14 of $\mathbf{w}^T \mathbf{x}$ to the pooled sample variance of $\mathbf{w}^T \mathbf{x}$. Maximisation of this ratio with
 15 respect to \mathbf{w} results in $\mathbf{w} = \alpha S^{-1}(\hat{\mu}_1 - \hat{\mu}_2)$, where α is an arbitrary scalar (not
 16 necessarily $n - 2$ as in [1]). Traditionally α is set to be 1 with the threshold c
 17 being adapted accordingly.

18 Fisher's linear discriminant rule does not assume Gaussian distributions for
 19 $\mathbf{x}|\omega_1$ and $\mathbf{x}|\omega_2$. However, in theory, it is equivalent to plug-in sample Gaussian-
 20 based LDA if the data satisfy the assumptions underlying the latter; in prac-
 21 tice, it can be equivalent to the latter with $c = -w_0$. However, when the

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

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

12 2 Comments on the Claim

13 We suggest that the above mentioned claim and the empirical study to support
14 it are vague if not misleading, even under an “ideal” condition such that $\hat{\mu}_j$
15 and S_j perfectly estimate μ_j and Σ_j , respectively. Let us explain it in the
16 context of three issues.

17 First, if the true prior probabilities are approximately balanced such that
18 $p(\omega_1) \approx p(\omega_2) \approx 0.5$ but the training set is unbalanced such that $n_1 \gg n_2$, then
19 plug-in estimates $\hat{p}(\omega_j)$ are poor estimates of $p(\omega_j)$ because $\hat{p}(\omega_1) \gg \hat{p}(\omega_2)$,
20 even though when the two sample covariance matrices are identical S will be
21 a good estimate of Σ . Consequently, being based on $\frac{\hat{p}(\omega_1)}{\hat{p}(\omega_2)}$, w_0 is wrongly esti-
22 mated so that LDA performs poorly. In this case, the use of re-balanced data,
23 as in [1], will no doubt adjust $\hat{p}(\omega_j)$ such that $\hat{p}(\omega_j) \approx 0.5$ and thus improve

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

5 Secondly, if the true prior probabilities are unbalanced such that $p(\omega_1) \gg$
 6 $p(\omega_2)$ and the training set demonstrates the imbalance such that $n_1 \gg n_2$,
 7 then plug-in estimates $\hat{p}(\omega_j) \approx p(\omega_j)$ are good estimates of $p(\omega_j)$ and thus $S =$
 8 $\hat{p}(\omega_1)S_1 + \hat{p}(\omega_2)S_2$ approaches the pooled population (within-class) covariance
 9 matrix $\Sigma = p(\omega_1)\Sigma_1 + p(\omega_2)\Sigma_2$. When the two sample covariance matrices are
 10 different, such that $S_1 \neq S_2$, the weights $\hat{p}(\omega_j)$ truly reflect the contribution
 11 of Σ_j to Σ . In contrast, if the training set is re-balanced by sampling as in [1],
 12 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
 13 no reason to suggest that an LDA that uses $S = \frac{1}{2}(S_1 + S_2)$ and a wrongly
 14 estimated w_0 (with the term $\log \frac{\hat{p}(\omega_1)}{\hat{p}(\omega_2)} = 0$) will perform better than LDA
 15 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
 16 Ref. [1] uses accurate estimates of the prior probabilities $\hat{p}(\omega_j)$ from the original
 17 data such that $\hat{p}(\omega_j) \approx p(\omega_j)$ and uses the re-balanced data to estimate the
 18 pooled covariance matrix such that $S = \frac{1}{2}(S_1 + S_2)$ for Gaussian-based LDA,
 19 there is still no justification that such a linear classifier will approach the
 20 performance of the best “admissible” linear procedure under the condition
 21 that $\Sigma_1 \neq \Sigma_2$ [3], which is similar to Fisher’s linear discriminant but with
 22 $\mathbf{w} = (t_1\Sigma_1 + t_2\Sigma_2)^{-1}(\mu_1 - \mu_2)$ (or in practice using sample statistics such that
 23 $\mathbf{w} = (t_1S_1 + t_2S_2)^{-1}(\hat{\mu}_1 - \hat{\mu}_2)$), where desired values of the scalars t_1 and t_2
 24 have no closed-form solution so that systematic trials or computing algorithms
 25 have to be adopted [3,4,5].

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

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

2 where $P(\omega_j|\omega_k)$ is the probability of misclassifying an observation, who arises
 3 from class ω_k , into class ω_j . For plug-in sample Gaussian-based LDA, when
 4 $(\mathbf{x}|\omega_1) \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $(\mathbf{x}|\omega_2) \sim \mathcal{N}(\mu_2, \Sigma_2)$, it follows that,

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

5

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

6 Similarly to [4], the estimated probabilities of misclassification can be rewrit-
 7 ten as

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

8

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

9 where Φ is the cumulative distribution function (CDF) of the standard nor-
 10 mal distribution $\mathcal{N}(0, 1)$. Therefore, in the formula for ER, $p(\omega_j)$ and Σ_j are
 11 population parameters, or sample parameters from a sufficiently large original
 12 dataset, while $\hat{p}(\omega_j)$, $\hat{\mu}_j$ and S are sample statistics obtained from a training
 13 set.

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

1 experiments, we assume that all the parameters of the linear discriminant
 2 function are estimated from the re-balanced training set such that $\hat{p}(\omega_j) \approx \frac{1}{2}$
 3 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
 4 re-balanced data can reduce ER can be translated into the following equality:

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

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

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

12 where

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

13

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

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

15 Here we consider a simple case in which each observation only has one feature
 16 (*i.e.*, $d = 1$). The population parameters are known to be $p(\omega_1) = 0.8$, $\mu_1 = 1$,
 17 $\mu_2 = -1$, $\Sigma_1 = 1$ and $\Sigma_2 \in [0.2, 5.0]$. The relationship between $\operatorname{ER}(\hat{p}(\omega_1))$ and
 18 $\hat{p}(\omega_1)$ is drawn in the 3-Dimensional plot as a function of $\hat{p}(\omega_1)$ and Σ_2 in the
 19 left panel of Figure 1. The surface of $\operatorname{ER}(\hat{p}(\omega_1))$ does not have a minimum

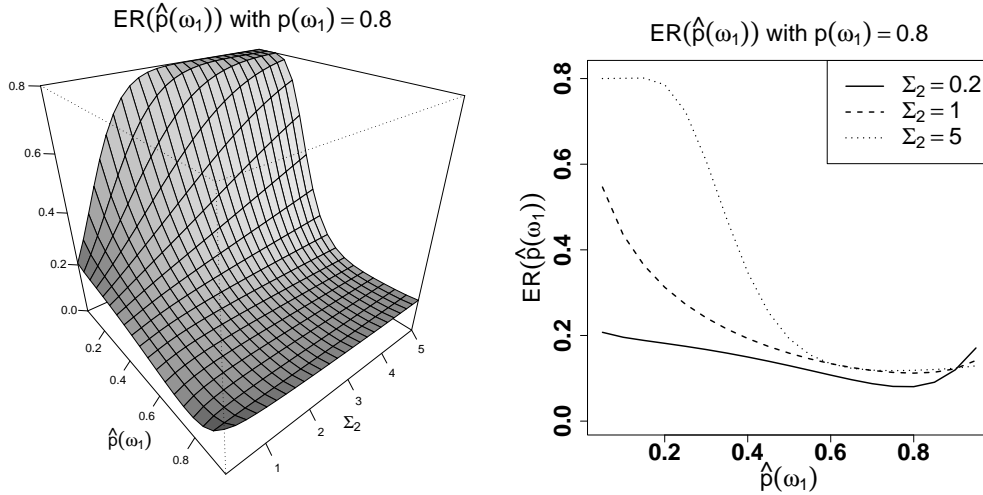


Figure 1. The misclassification error rates $ER(\hat{p}(\omega_1))$.

- 1 point at $\hat{p}(\omega_1) = 0.5$.
- 2 In the right panel of Figure 1, we draw the curves of $ER(\hat{p}(\omega_1))$ for $\Sigma_2 = 0.2, 1,$
- 3 and 5, respectively. We observe the following.
 - 4 (1) When $\Sigma_2 = 0.2$ or 5 such that $\Sigma_2 \neq \Sigma_1$, the best performance of LDA is
 - 5 obtained at $\hat{p}(\omega_1) = 0.8$, which is equal to the true prior probability of
 - 6 class ω_1 , rather than from the re-balanced data, which give $\hat{p}(\omega_1) = 0.5$;
 - 7 the procedure of re-balancing data has a negative effect on the perfor-
 - 8 mance of LDA if the original unbalanced data conform to the truly un-
 - 9 balanced population.
 - 10 (2) When $\Sigma_2 = 1$ such that $\Sigma_2 = \Sigma_1$, the best performance of LDA is also
 - 11 obtained at $\hat{p}(\omega_1) = 0.8$ rather than from the re-balanced data; the pro-
 - 12 cedure of re-balancing data may also have a negative effect.
 - 13 (3) In general, data with a compact within-class distribution (in the sense of a
 - 14 small within-class covariance matrix) may result in a better performance
 - 15 of LDA (in the sense of smaller $ER(\hat{p}(\omega_1))$), compared with data with a
 - 16 dispersed within-class distribution.

1 (4) In fact, in this case, since $\min(p(\omega_1), p(\omega_2)) = 0.2$, in practice the maxi-
2 mum $\text{ER}(\hat{p}(\omega_1))$ can be controlled to be 0.2, the smaller prior probability,
3 if we always classify observations into the class with higher prior proba-
4 bility.

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

11 **3 AUC or ER**

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

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

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

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

10 For two-class discrimination of unbalanced data, where the error in the mi-
11 nority class may be more important in practice, the Receiver Operating Char-
12 acteristic (ROC) curve and the area under the curve, the so-called AUC, are
13 commonly used [7,9]. The ROC curve is a plot of the true positive rate vs.
14 the false positive rate, and hence a higher AUC generally indicates a better
15 classifier. As pointed out by [10], there is a three-way equivalence between
16 AUC, the Wilcoxon-Mann-Whitney statistic and the probability of a correct
17 ranking of a randomly chosen (negative, positive) pair. More precisely, sup-
18 pose that a discriminant function such as $g^L(\mathbf{x})$ is designed to provide a high
19 score for a positive observation and a low score for a negative one. Then, given
20 a randomly chosen (negative, positive) pair denoted by $(\mathbf{x}_N, \mathbf{x}_P)$, it holds that
21 $AUC = Prob(\mathbf{x}_N < \mathbf{x}_P)$.

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

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

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

18 Ref. [1] uses AUC to evaluate the performance of plug-in sample Gaussian
19 LDA (denoted by LDA- Σ hereafter); in our study, we will use both AUC and
20 ER to evaluate the performance of LDA- Σ and one of its special versions which
21 assumes that the common covariance matrix is diagonal (denoted by LDA- Λ).
22 In our implementation, we first carry out experiments on 15 unbalanced (with
23 the proportion of the majority class $\hat{p}(\omega_2) > 65\%$) datasets. Obtained from
24 the UCI machine learning repository [13], the datasets include all 10 datasets
25 used by [1] and 5 other more unbalanced datasets (with $\hat{p}(\omega_2) > 75\%$); as
26 with [1], these datasets have only continuous features. Then, we investigate 4

1 simulated datasets of normally distributed data and normal mixture data.

2 4 Replication of Experiments on UCI Datasets

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

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

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

21 Table 1 presents the description of the 10 UCI datasets being studied by
22 both [1] and us (the class prior probabilities different from Table 1 of [1] are
23 highlighted in italics). The experiments on the 5 other UCI datasets provide

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

1 similar results which can be found in the appendix of a report on the web
2 page for Technical Reports of the Department of Statistics at the University
3 of Glasgow.

4 As with [8] and [14], the UCI data are rescaled into the range $[0, 1]$. In addition,
5 before carrying out LDA, we perform, for each feature $\mathbf{x}_i|\mathbf{y}$, the Shapiro-Wilk
6 test for within-class normality and Levene’s test for homogeneity of variance
7 across the two classes at the significance level 0.05. As the maximum number

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.

1 of observations allowed by an R function *shapiro.test* from the R package **stats**
2 is 5000, we use 5000 randomly sampled observations for the tests when there
3 are more in the dataset. If for a particular feature the within-class normality
4 is rejected in either of the two classes, we mark the feature as “Normality
5 rejected”. Results of these two tests, as shown in Table 2, suggest that for
6 all 10 datasets under study the null hypotheses of within-class normality and
7 homoscedasticity across the classes are rejected, including the dataset “Pima”

1 which is stated to have nearly equal sample covariance matrices in [1].

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).

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

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).

- 1 (1) Concerning LDA- Σ , AUCs of re-balanced data are significantly (at the
- 2 level 0.05) better than those of original data, except for the under-sampled
- 3 data of “Pima”, “New-thyroid” and “Glass”. Although the increase of its
- 4 median (and thus the improvement of classifier performance) from re-
- 5 balancing is not very large in amount, in general, it can be said that, for
- 6 the datasets being studied, AUC favours re-balanced data.

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).

- 1 (2) Concerning LDA- Λ : of the 10 datasets, AUCs of re-balanced “Satimage-
2 3” and “Image” are significantly worse than those of the original data
3 for both re-sampling methods, and AUC of re-balanced “Glass” is signif-
4 icantly worse than that of original data for the under-sampling. Mean-
5 while, no significant difference exists between AUCs of “Vehicle”. This
6 may be because of the different estimates of the covariance matrix be-

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).

1 tween LDA- Σ and LDA- Λ ; this indicates that the accuracy of estimation
2 can play a more important role in AUC than the re-balancing does.
3 (3) In contrast to AUC, ER is significantly increased by re-balancing except
4 for “New-thyroid” and “Vehicle”. The increase of its median (and thus
5 the decline of classifier performance) from re-balancing is relatively large.
6 In general, it can be said that, for the datasets being studied, ER favours

1 original data.

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

7 5 Simulation Studies

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

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

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

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

20 where population prior probabilities $p(\omega_j)$ are used for the term $\log \frac{p(\omega_1)}{p(\omega_2)}$ in w_0
21 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

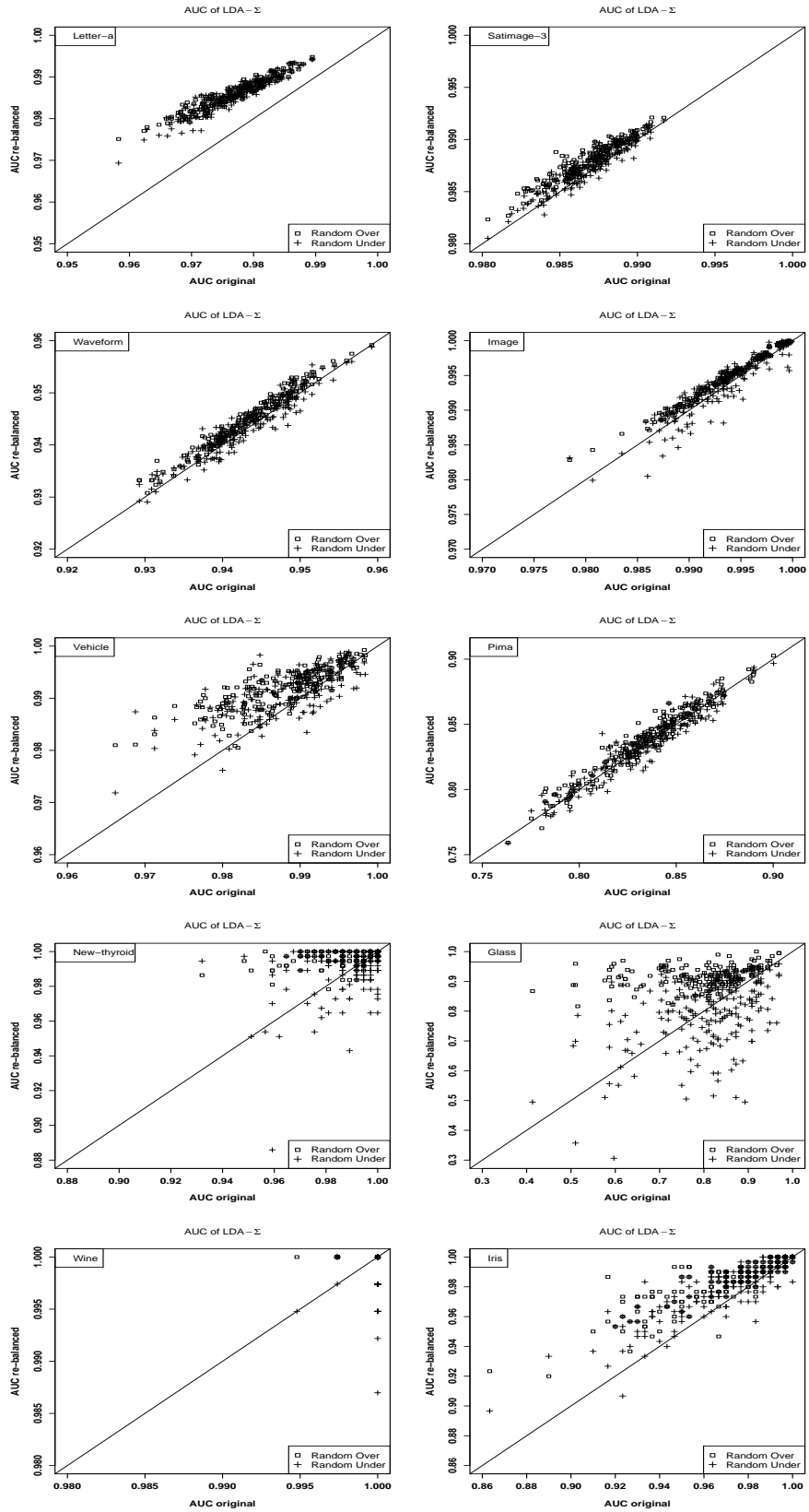


Figure 2. Scatter plots of AUC on re-balanced data (by over-sampling and under-sampling) vs. original data, obtained from LDA- Σ .

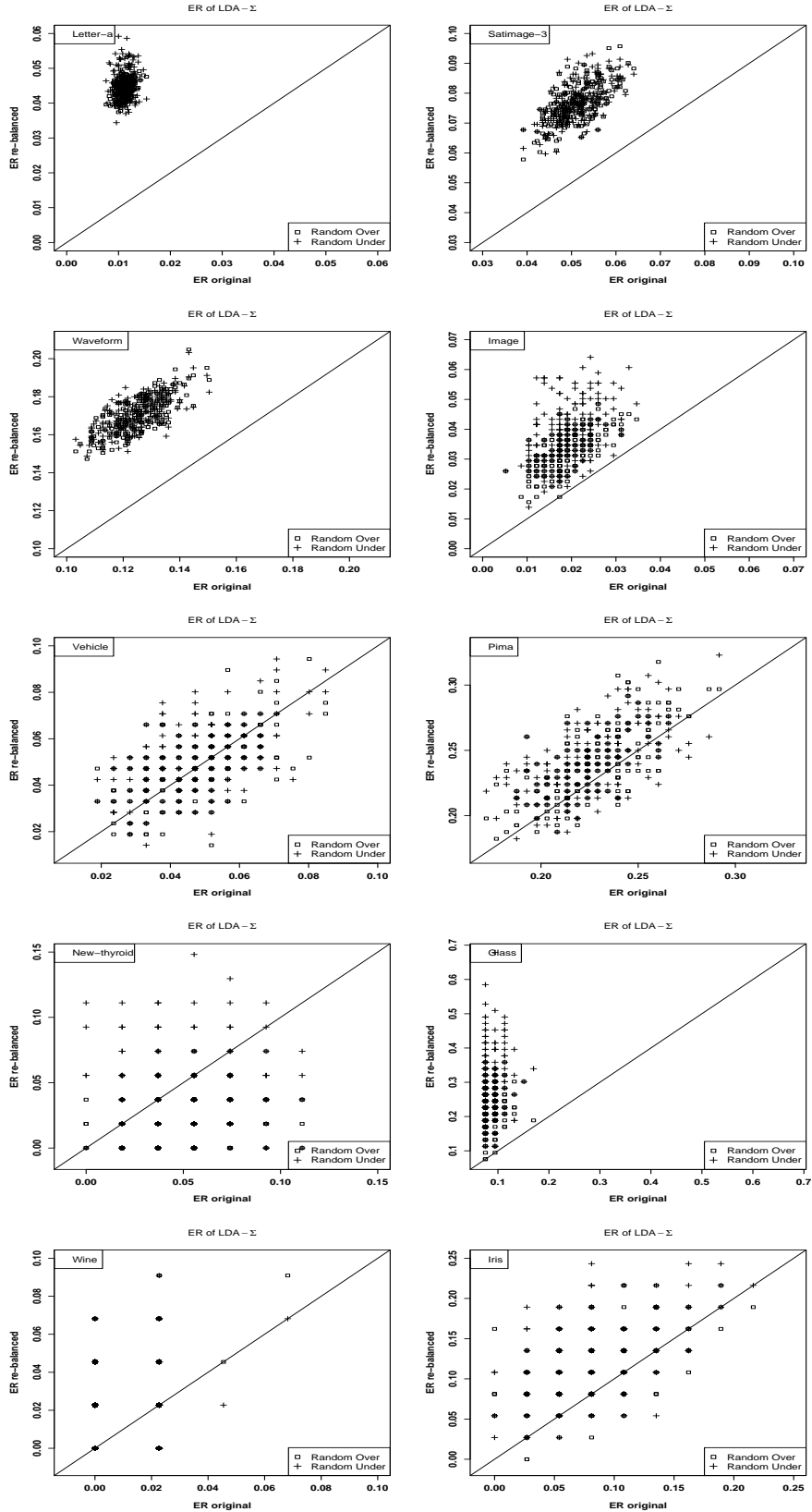


Figure 3. Scatter plots of ER on re-balanced data (by over-sampling and under-sampling) vs. original data, obtained from LDA- Σ .

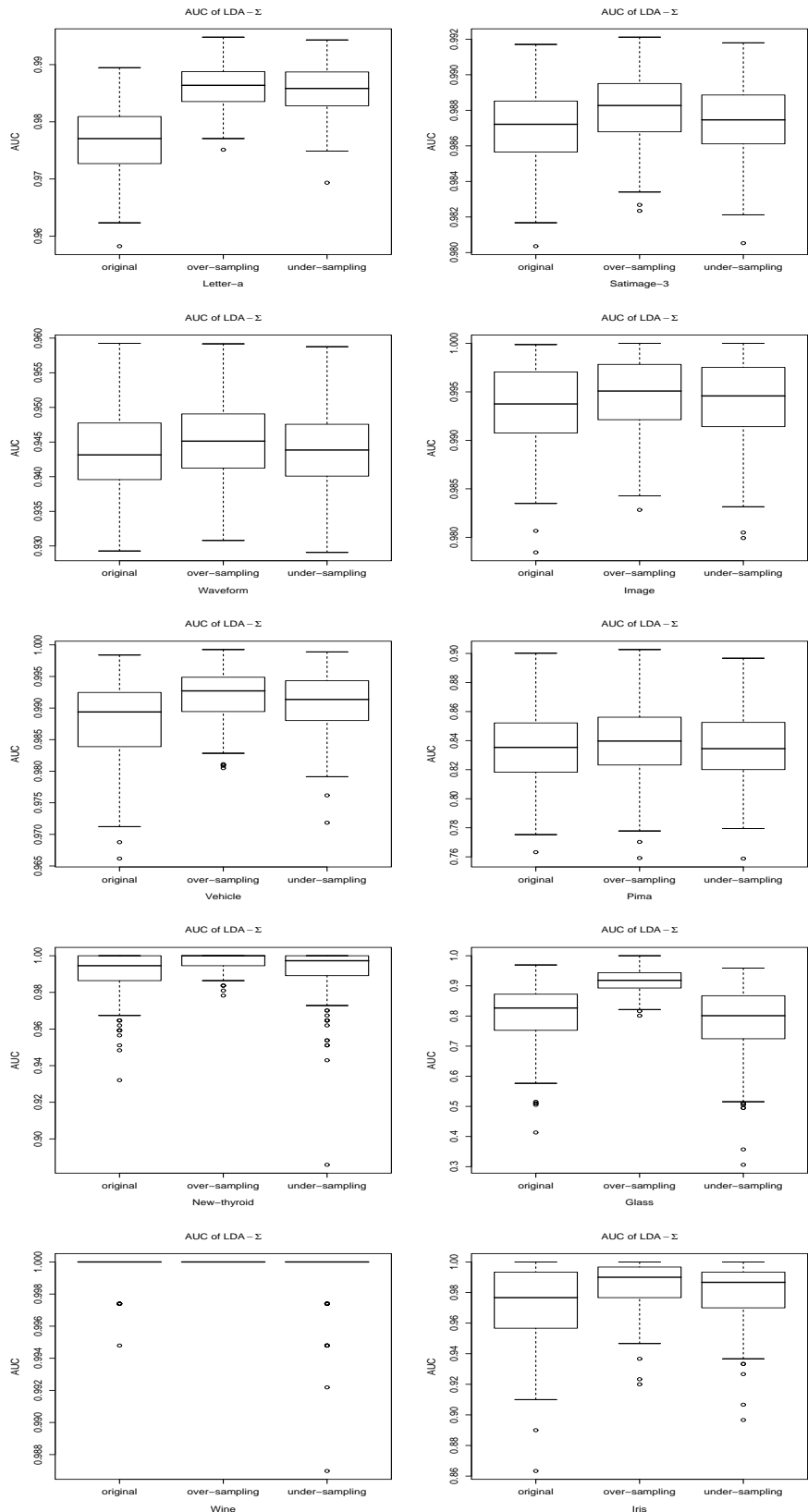


Figure 4. Box-plots of AUC on original and re-balanced data (by over-sampling and under-sampling), obtained from LDA- Σ .

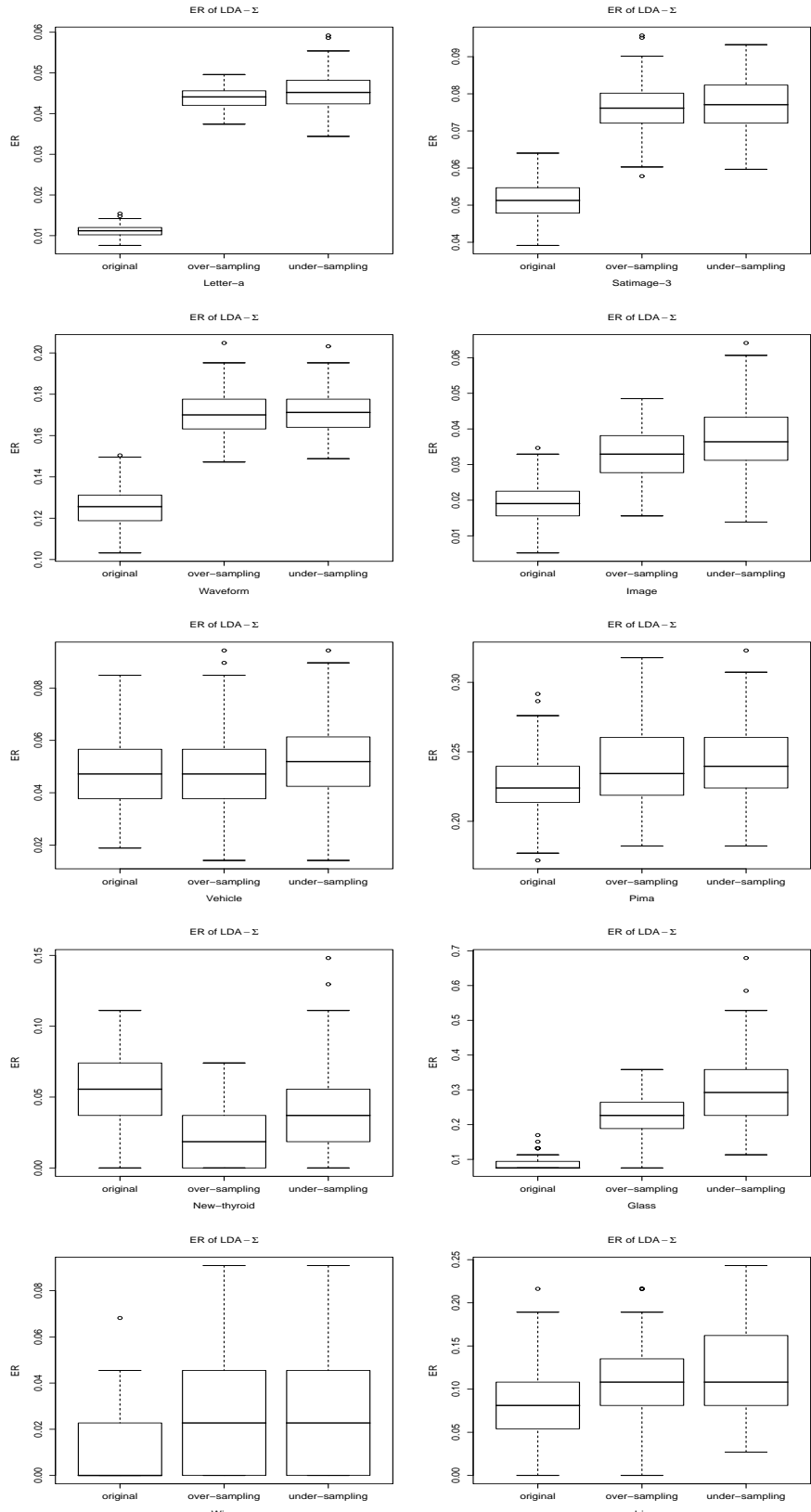


Figure 5. Box-plots of ER on original and re-balanced data (by over-sampling and under-sampling), obtained from LDA- Σ .

1 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 .

2 In this section, we simulate 4 datasets; each dataset consists of 1000 observa-
3 tions and is divided into two classes, ω_1 and ω_2 , with 200 observations from
4 the minority class ω_1 and 800 observations from the majority class ω_2 such
5 that each dataset is unbalanced with $\hat{p}(\omega_1) = 0.2$. The first dataset is ran-
6 domly generated from two 4-variate normal distributions, $\mathbf{x}|\omega_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$
7 and $\mathbf{x}|\omega_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$, with equal covariance matrices such that $\Sigma_1 = \Sigma_2$;
8 the second dataset is similar to the first one except that $\Sigma_1 \neq \Sigma_2$. The third
9 and fourth datasets are randomly generated from two 4-variate normal mix-
10 tures; each mixture has two components. The third one has equal covariance
11 matrices across the two classes while the fourth one does not.

12 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
13 linear transformation to reduce Σ_1 to the identity matrix \mathbf{I} and diagonalise Σ_2 .
14 Therefore, without loss of generality, in this section, we use a canonical form
15 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
16 covariance matrix. For the dataset with equal covariance matrices, $\Sigma_2 = \mathbf{I} =$
17 Σ_1 ; for the dataset with unequal covariance matrices, Σ_2 is a diagonal matrix
18 with 4 diagonal elements which are $(0.25, 0.75, 1.25, 1.75)$, so that $\Sigma_2 \neq \Sigma_1$.

19 Compared with the normal distribution, the mixture of normal distributions
20 is a better approximation to real data in a variety of situations. In this section,
21 2 simulated datasets are randomly generated from two mixtures, ω_1 and ω_2 ,
22 of 4-variate normal distributions.

23 Each mixture has two components with equal mixing coefficients. The two
24 components, A and B , of the mixture ω_1 are normally distributed with proba-

1 bility density functions $\mathcal{N}(\mu_{1A}, \Sigma_1)$ and $\mathcal{N}(\mu_{1B}, \Sigma_1)$, respectively, where $\mu_{1A} =$
2 $\mathbf{0}$ and $\mu_{1B} = (2, 0, 0, 0)^T$; and the two components, C and D , of the mixture
3 ω_2 are normally distributed with probability density functions $\mathcal{N}(\mu_{2C}, \Sigma_2)$ and
4 $\mathcal{N}(\mu_{2D}, \Sigma_2)$, respectively, where $\mu_{2C} = (-1.5, -0.75, 0.75, 1.5)^T$ and $\mu_{2D} =$
5 $(-3.5, -0.75, 0.75, 1.5)^T$. In such a way, when Σ_1 and Σ_2 are equal/unequal,
6 the covariance matrices of the two mixtures will become equal/unequal. Mean-
7 while, we set Σ_1 and Σ_2 in the same way as for the normally distributed data.

8 In our simulation studies, both Σ_1 and Σ_2 are diagonal; the performance of
9 LDA- Λ is similar to that of LDA- Σ , and thus only the results obtained from
10 LDA- Σ are presented in the following.

11 The simulations from the multivariate normal distributions and normal mix-
12 tures are based on an R function *mvnorm* for simulating, from a contributed
13 R package **MASS**. As with the UCI datasets being studied, the simulated
14 data are rescaled into the range $[0, 1]$.

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

19 (1) Concerning AUC obtained from LDA- Σ , although for the simulated datasets
20 being studied it generally favours re-balanced data, the increase of its
21 median (and thus the improvement of performance of LDA) from re-
22 balancing is relatively small. We observe that, of the two simulated mix-
23 ture datasets, there is no significant change in AUC between under-
24 sampled and original data for one dataset and between over-sampled and

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).

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).

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

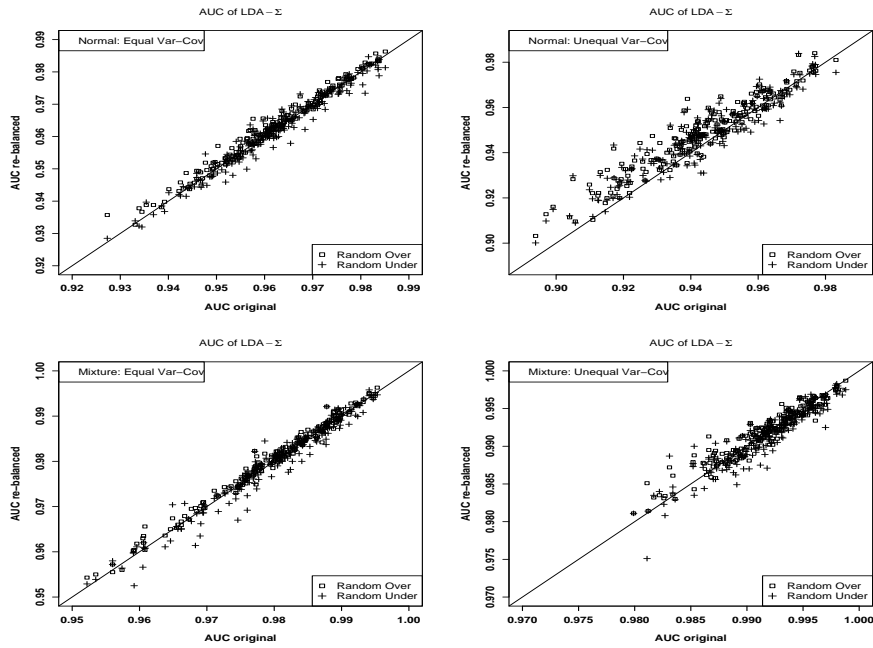


Figure 6. Scatter plots of AUC on re-balanced data (by over-sampling and under-sampling) vs. original data, obtained from LDA- Σ .

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

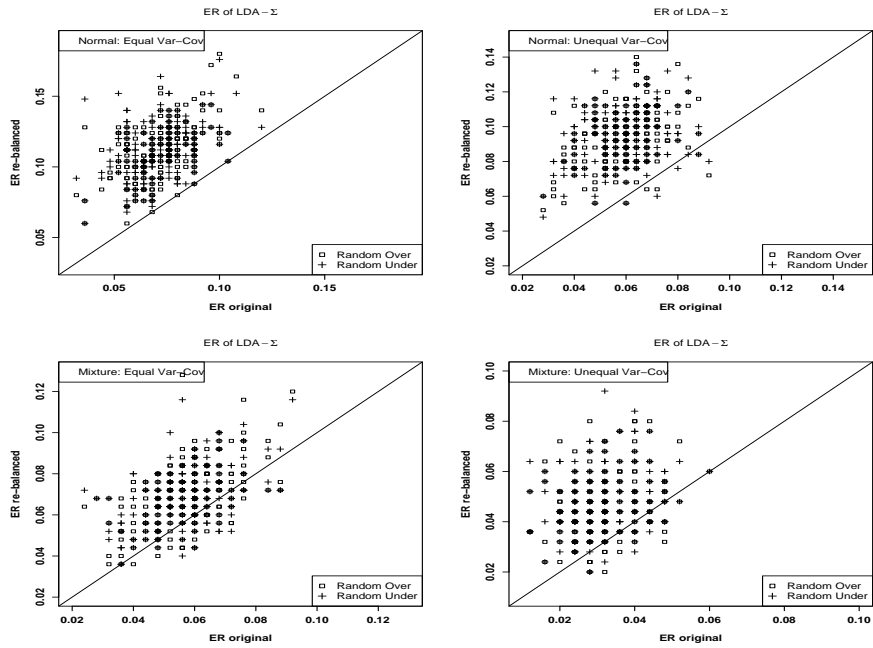


Figure 7. Scatter plots of ER on re-balanced data (by over-sampling and under-sampling) vs. original data, obtained from LDA- Σ .

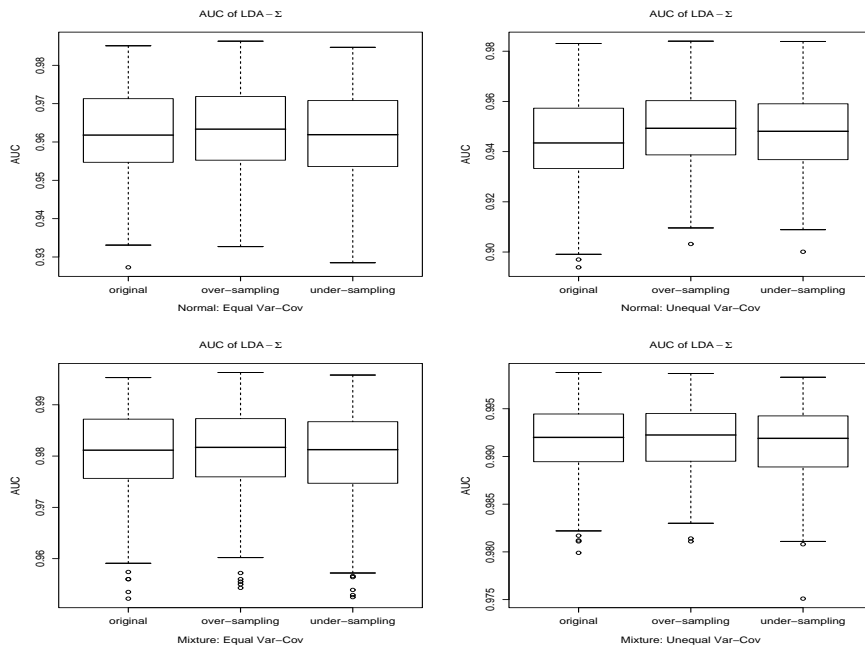


Figure 8. Box-plots of AUC on original and re-balanced data (by over-sampling and under-sampling), obtained from LDA- Σ .

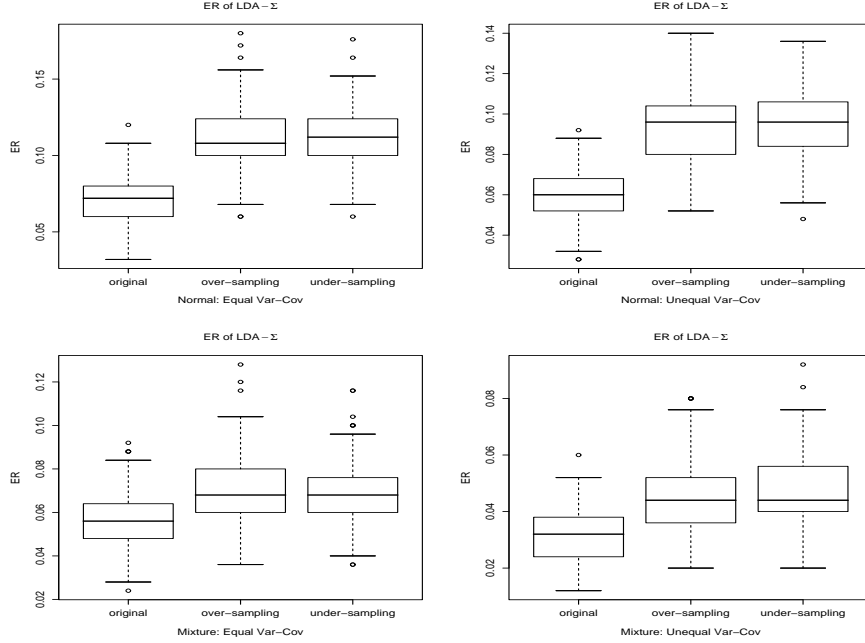


Figure 9. Box-plots of ER on original and re-balanced data (by over-sampling and under-sampling), obtained from LDA- Σ .

1 6 Conclusions

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

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

11 (2) Therefore, from our study, there is no reliable empirical evidence to sup-
12 port the claim that a (class) unbalanced data set has a negative effect on

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

6 **References**

- 7 [1] J. G. Xie, Z. D. Qiu, The effect of imbalanced data sets on LDA: a theoretical
8 and empirical analysis, *Pattern Recognition* 40 (2) (2007) 557–562.
- 9 [2] E. S. Gilbert, The effect of unequal variance-covariance matrices on Fisher’s
10 linear discriminant function, *Biometrics* 25 (3) (1969) 505–515.
- 11 [3] T. W. Anderson, R. R. Bahadur, Classification into two multivariate normal
12 distributions with different covariance matrices, *The Annals of Mathematical*
13 *Statistics* 33 (2) (1962) 420–431.
- 14 [4] S. Marks, O. J. Dunn, Discriminant function when covariance matrices are
15 unequal, *Journal of the American Statistical Association* 69 (345) (1974) 555–
16 559.
- 17 [5] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*,
18 John Wiley & Sons, New York, 1992.
- 19 [6] D. M. Titterington, G. D. Murray, L. S. Murray, D. J. Spiegelhalter, A. M.
20 Skene, J. D. F. Habbema, G. J. Gelpke, Comparison of discrimination
21 techniques applied to a complex data set of head injured patients (with
22 discussion), *Journal of the Royal Statistical Society. Series A (General)* 144 (2)
23 (1981) 145–175.

- 1 [7] G. M. Weiss, Mining with rarity: a unifying framework, SIGKDD Explorations
2 6 (1) (2004) 7–19.
- 3 [8] A. P. Bradley, The use of the area under the ROC curve in the evaluation of
4 machine learning algorithms, Pattern Recognition 30 (7) (1997) 1145–1159.
- 5 [9] N. V. Chawla, N. Japkowicz, A. Kotcz, Editorial: special issue on learning from
6 imbalanced data sets, SIGKDD Explorations 6 (1) (2004) 1–6.
- 7 [10] J. A. Hanley, B. J. McNeil, The meaning and use of the area under a receiver
8 operating characteristic (ROC) curve, Radiology 143 (1) (1982) 29–36.
- 9 [11] C. Cortes, M. Mohri, AUC optimization vs. error rate minimization, in: NIPS,
10 2003.
- 11 [12] G. M. Weiss, F. Provost, Learning when training data are costly: the effect of
12 class distribution on tree induction, Journal of Artificial Intelligence Research
13 19 (2003) 315–354.
- 14 [13] D. J. Newman, S. Hettich, C. L. Blake, C. J. Merz, UCI Repository of machine
15 learning databases, University of California, Irvine, Dept. of Information
16 and Computer Sciences, <http://www.ics.uci.edu/~mlearn/MLRepository.html>
17 (1998).
- 18 [14] A. Y. Ng, M. I. Jordan, On discriminative vs. generative classifiers: a comparison
19 of logistic regression and naïve bayes, in: NIPS, 2001, pp. 841–848.

20 **7 Appendix: Results on 5 UCI datasets**

Data set	Observations	Features	Class (min., maj.)	Prior (min., maj.)
Optdigit-0	5,620	62	(0, remainder)	(9.86%, 90.14%)
Pendigit-0	10,992	16	(0, remainder)	(10.4%, 89.6%)
Wpbc	198	30	(R, N)	(23.74%, 76.26%)
Shuttle	14,500	9	(remainder, 1)	(20.84%, 79.16%)
Vowel-context	990	10	(0, remainder)	(9.09%, 90.91%)

Table 9

Description of data

Data set	Features	Normality rejected	Homoscedasticity rejected
Optdigit-0	62	62	49
Pendigit-0	16	16	13
Wpbc	30	27	1
Shuttle	9	9	7
Vowel-context	10	10	9

Table 10

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

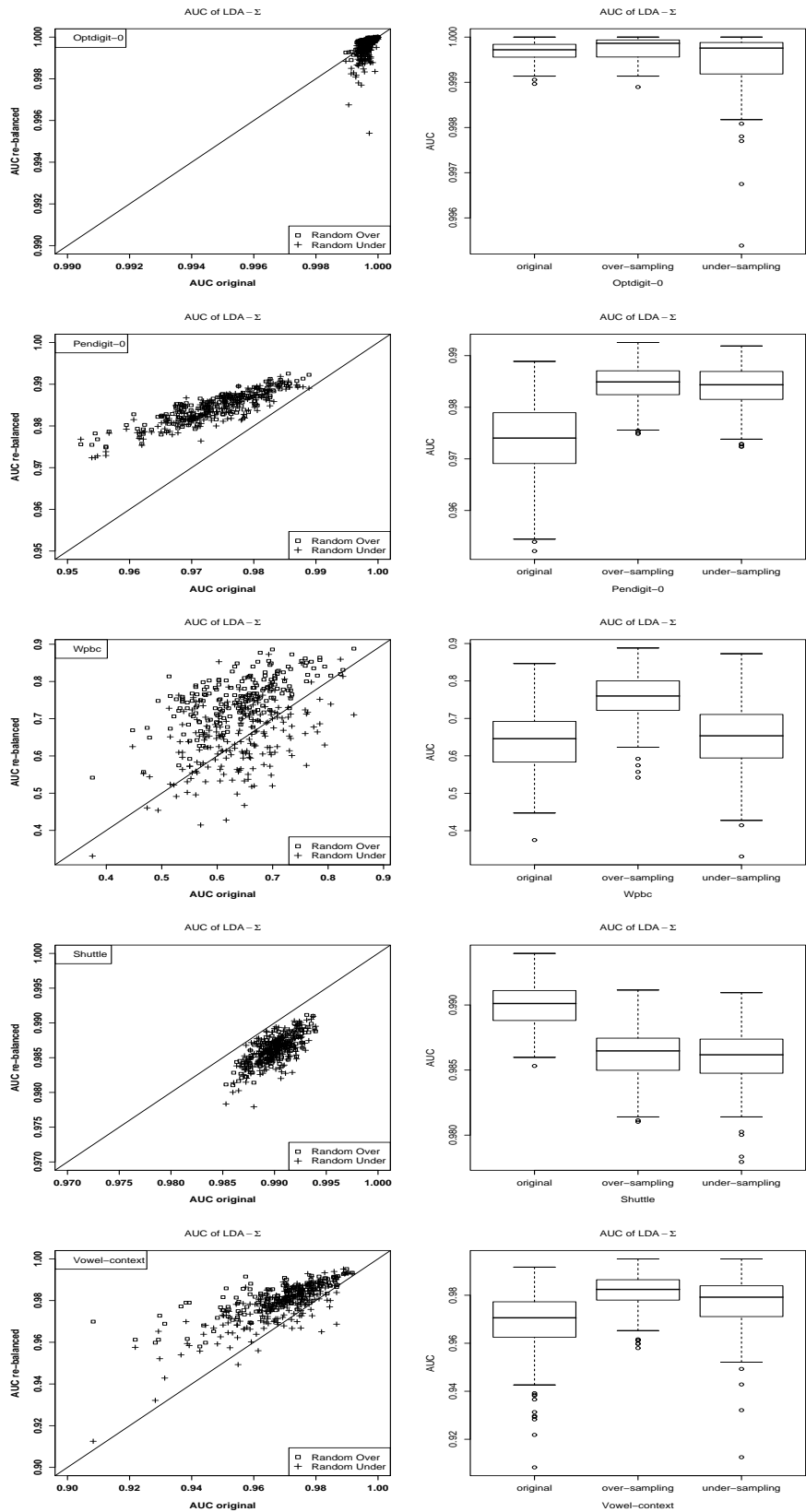


Figure 10. Scatter plots of AUC on re-balanced data (by over-sampling and under-sampling) vs. original data and corresponding box-plots of AUC on original and re-balanced data, obtained from LDA- Σ .

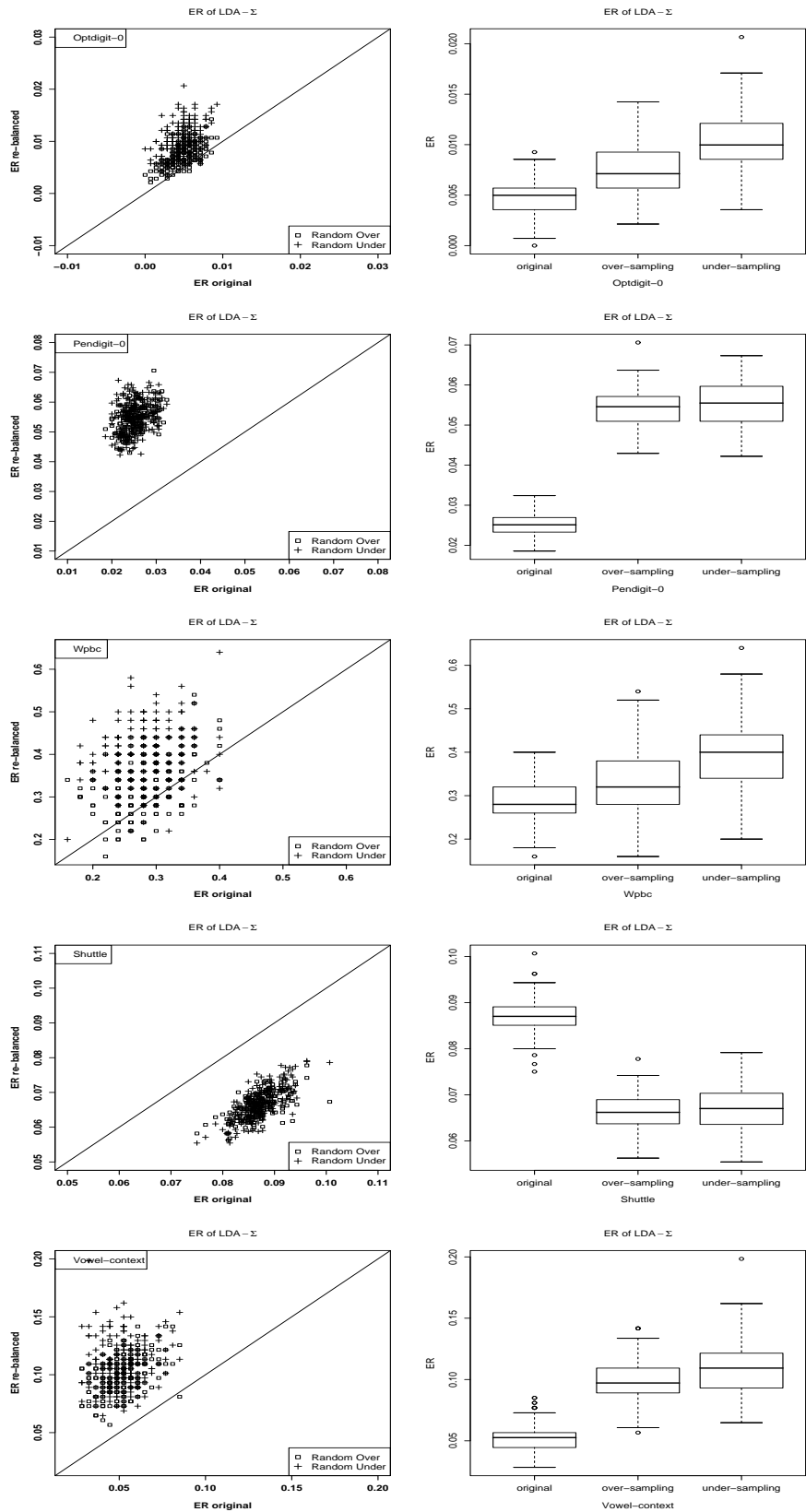


Figure 11. Scatter plots of ER on re-balanced data (by over-sampling and under-sampling) vs. original data and corresponding box-plots of ER on original and re-balanced data, obtained from LDA- Σ .

Data set	Original	Over.	Under.	p-v. (Ori.-Over.)	p-v. (Ori.-Under.)
Optdigit-0	1	1	1	0	0.001
Pendigit-0	0.974	0.985	0.984	0	0
Wpbc	0.646	0.760	0.654	0	0.283
Shuttle	0.990	0.986	0.986	0	0
Vowel-context	0.971	0.982	0.979	0	0

Table 11

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).

Data set	Original	Over.	Under.	p-v. (Ori.-Over.)	p-v. (Ori.-Under.)
Optdigit-0	0.005	0.007	0.010	0	0
Pendigit-0	0.025	0.055	0.055	0	0
Wpbc	0.280	0.320	0.400	0	0
Shuttle	0.087	0.066	0.067	0	0
Vowel-context	0.053	0.097	0.109	0	0

Table 12

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).

Data set	Original	Over.	Under.	p-v. (Ori.-Over.)	p-v. (Ori.-Under.)
Optdigit-0	0.999	0.999	0.999	0	0
Pendigit-0	0.968	0.973	0.973	0	0
Wpbc	0.602	0.626	0.596	0	0.843
Shuttle	0.974	0.976	0.976	0	0
Vowel-context	0.957	0.964	0.961	0	0

Table 13

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).

Data set	Original	Over.	Under.	p-v. (Ori.-Over.)	p-v. (Ori.-Under.)
Optdigit-0	0.016	0.028	0.028	0	0
Pendigit-0	0.095	0.113	0.113	0	0
Wpbc	0.360	0.400	0.440	0	0
Shuttle	0.077	0.076	0.075	0	0
Vowel-context	0.069	0.121	0.121	0	0

Table 14

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).