NORMAL APPROXIMATION USING BOX-COX TRANSFORMATIONS IN DISCRIMINANT ANALYSIS FOR THE TWO-CLASS PROBLEM

Roger P. Qu, PhD
Department of Biostatistics

MADISON, WISCONSIN
Normal Approximation Using Box-Cox
Transformations in Discriminant Analysis for the
Two-Class Problem

Roger P. Qu

Department of Biotatistics, University of Wisconsin-Madison,
K6/446 Clinical Sciences Center, 600 Highland Avenue, Madison, WI 53792

Key words and Phrases: linear discriminant; quadratic discriminant; unequal variances.

ABSTRACT

Normal approximation using Box-Cox transformations in discriminant analysis for two classes with
or without negative observations is examined. The transformation parameter is estimated assum-
ing normal approximations with equal and unequal variances. Linear and quadratic discriminant
analysis applied to the transformed data are studied and the limiting error probabilities with and
without transformation are obtained. Numerical comparison of the limiting error probabilities
with simulated finite sample probabilities are given. It is found that the normal approximation can
reduce the error probabilities significantly.
1 Introduction

The purpose of this paper is to study the use of normal approximations via Box-Cox transformations (1964) in discriminant analysis. In a two-class discriminant problem with populations 1 and 2, one wants to classify an observation \( Z \) to one of these two populations. If the two populations are totally or partially unknown, training samples, one from each population, are needed. The usual assumption in discriminant analysis is that the underlying distributions are normal. Under this assumption, the maximum likelihood rule (MLR), which gives the smallest total misclassification error probability, is the Gaussian linear discriminant rule (LDR) if the variances are equal and the quadratic discriminant rule (QDR) otherwise. If the parent distributions are not normal, then neither LDR nor QDR is necessarily the best. Many studies have examined the consequences of violation of the normality assumption. It is known that LDR and QDR can be far from optimal if the underlying distributions depart greatly from normal. One possible way of overcoming this problem is to transform the data to asymptotic normality using Box-Cox transformations before applying LDR or QDR.

Lachenbruch, Sneeringer and Revo (1973) studied the effect of using LDR on log normal, logit normal and inverse hyperbolic sine normal distributions by comparing the error probabilities before and after transformations. Their work was under the assumption that the transformed data, using log, logit or inverse hyperbolic transformations, have exact normal distributions with equal variances. They found that the performance of LDR could be severely affected by non-normality of the parent distributions. Beauchamp, Folkert and Robson (1980) and Beauchamp and Robson (1986) investigated the effect of LDR and QDR in the case of non-normality by comparing the error probabilities before and after Box-Cox transformations for distributions that are exact normal with equal covariance matrices after transformations.
much of the previous results were obtained under the assumption that the transformed data have exact normal distributions with equal covariance matrices. In most practical cases, however, the underlying distributions cannot be transformed to exact normal distributions. Thus it is natural to investigate the effect of transformations on LDR and QDR from the other direction, i.e., to assume that the transformed distributions are not exact normal but that the untransformed distributions are of some common types of distributions such as the gamma, exponential or lognormal distributions. Under this assumption, Qu and Loh (1992) investigated the effect of transforming the underlying non-normal distributions to approximate symmetric distributions using Hinkley's method (1975). It was found that the misclassification probabilities can be reduced significantly for the lognormal, gamma, exponential and log-double exponential distributions.

Another characteristic of previous work on the performance of transformations is that the distributions and transformation parameter in Box-Cox transformations were totally known. In practical situations, the population distributions are usually unknown. This is why sample means and variances are used in the LDR or QDR methods. Therefore meaningful comparisons of error probabilities should start at this point. Also, if the true population distributions are unknown, the transformation parameter is unknown too no matter what criterion is used in the Box-Cox transformations. Therefore one has to estimate the transformation parameter using training samples. In the present work, it is assumed that the two parent distributions are unknown but differ only by an unknown location shift. The performance of the normal approximation will be investigated intensively for univariate response in the first 5 sections. The effect of using LDR and QDR on the transformed data will be considered and compared to LDR on the original data. Section 6 summarizes results for multivariate case.
The family of Box-Cox transformations transform an observed variable $X$ to $X_\lambda$:

$$X_\lambda = \begin{cases} 
(X^\lambda - 1)/\lambda, & \text{if } \lambda \neq 0, \\
\log(X), & \text{otherwise.}
\end{cases}$$

There are two possible ways of using Box-Cox transformations to achieve approximate normality in discriminant analysis: with equal variances and with unequal variances. Notice that the Box-Cox transformation is applicable only to positive random variables. It is not uncommon in practice to have data with negative observations. Sections 2, 3 and 4 deal with positive observations only. Section 5 considers data with negative observations.

2 Normal approximation using Box-Cox transformations

Let $X_1, X_2, \ldots, X_n$, be i.i.d. positive random variables from population 1 having c.d.f. $F(x)$, and $Y_1, Y_2, \ldots, Y_n$ be i.i.d. from population 2 having c.d.f. $F(x-c)$, where $c > 0$ is an unknown location shift. Given a further observation $Z$ either from population 1 or 2, we want to classify $Z$ to one of the two populations.

Let $\bar{X} = n_1^{-1} \sum_{i=1}^{n_1} X_i$ and $\bar{Y} = n_2^{-1} \sum_{i=1}^{n_2} Y_i$ be the sample means of the $X$'s and $Y$'s respectively. The Gaussian linear discriminant function (LDF) is $W = Z - (\bar{X} + \bar{Y})/2$ which assigns $Z$ to population 2 if $W$ is non-negative and to population 1 otherwise. The two types of misclassification probabilities are $P[W \geq 0|Z \sim F(x)]$ and $P[W < 0|Z \sim F(x-c)]$.

In the following we assume that random variable $X$ has c.d.f. $F(x)$ with mean $\mu$ unless otherwise specified. By the Law of Large Numbers, the limiting distribution of $W$ is that of $Z - (\mu + c/2)$. Therefore the total error probability converges to

$$p = P(|X - \mu| > c/2).$$
(1)
Let $\bar{X}_\lambda$ be the mean of $X_{\lambda,1}, X_{\lambda,2}, \ldots, X_{\lambda,n_1}$, the transformed data of $X_1, X_2, \ldots, X_{n_1}$, i.e.

$$
\bar{X}_\lambda = \begin{cases} 
  n_1^{-1} \sum_{i=1}^{n_1} (X_i^\lambda - 1)/\lambda, & \text{if } \lambda \neq 0, \\
  n_1^{-1} \sum_{i=1}^{n_1} \log(X_i), & \text{otherwise,}
\end{cases}
$$

and similarly for $\bar{Y}_\lambda$. Let $n = n_1 + n_2$ be the total sample size. Suppose that $n_1/n_2 \to k \in (0, \infty)$ for some constant $k$ as $n_1$ and $n_2$ increase to infinity.

Let $\hat{\lambda}$ denote an estimate of $\lambda$ according to some criterion and based on the two samples $X_1, X_2, \ldots, X_{n_1}$, and $Y_1, Y_2, \ldots, Y_{n_2}$. The dependence of $\hat{\lambda}$ on the sample size $n = n_1 + n_2$ will be suppressed. Throughout this chapter, we make the following assumptions unless otherwise specified:

A1. There are constants $a$ and $b$ such that $E(X_\lambda^2) < \infty$ for $\lambda \in [a, b]$.

A2. $E|\log(X)| < \infty$.

A3. $\hat{\lambda} \in [a, b]$.

Lemma 1 Conditions A1 and A2 imply the same with random variable $X$ replaced by $X + c$ for any $c > 0$.

Proof. First, we see that $EX_\lambda^2 < \infty$ is equivalent to $EX^{2\lambda} < \infty$ if $\lambda \neq 0$ and to $E[|\log(X)|]^2 < \infty$ if $\lambda = 0$. For any constant $\lambda > 0$, there is a constant $d_\lambda$ depending on $\lambda$ only such that $|(X + c)\lambda| \leq [d_\lambda(X^\lambda + c^\lambda) + 1]/\lambda$, and for $\lambda < 0$, $|(X + c)\lambda| \leq [X^\lambda + 1]/|\lambda|$. So $EX^{2\lambda} < \infty$ implies $E(X + c)^{2\lambda} < \infty$ for $\lambda \neq 0$. To see that $E[|\log(X)|]^2 < \infty$ implies that $E[|\log(X + c)|]^2 < \infty$, let $x_0 \geq 1$ be such that $x_0 + c < x_0e^c$, which is equivalent to $1 + c/x_0 < e^c$. Then

$$
E[|\log(X + c)|]^2 \leq E[|\log(X + c)|I(X \geq x_0) + |\log(X + c)|I(X \leq x_0)]^2
\leq E[|\log(X + c)|I(X \geq x_0) + |\log(c)| + \log(x_0 + c)]^2
\leq E[|\log(X + c)|I(X \geq x_0)]^2 + [|\log(c)| + \log(x_0 + c)]^2
$$
+ 2E \log(X + c) I(X \geq x_0)\log(c) + \log(x_0 + c),

where \( I(A) \) is the indicator function of a set \( A \). Consider the function \( f(x) = \log(x + c) - \log(x) - c \), for \( x \geq x_0 \). Then \( f(x_0) < 0 \) and \( f'(x) < 0 \) for \( x \geq x_0 \). Thus \( f(x) < 0 \) for \( x \geq x_0 \). Therefore \( \log(x + c)I(x \geq x_0) \leq \log(x)I(x \geq x_0) + c \) and

\[
E[\log(X + c)I(X \geq x_0)]^2 \leq E[\log(X)I(X \geq x_0) + c]^2 \\
\leq E[\log(X)]^2 + 2cE[\log(X)] + c^2.
\]

Hence \( E[\log(X + c)]^2 < \infty \).

Note that if \( a < 0 < b \), then A2 is implied by A1 but not otherwise.

2.1 Normal approximation with equal variances

This is the most common way of using Box-Cox transformations. It is assumed that there is a \( \lambda_e \) such that the transformed distributions of \( X_{\lambda_e} \) and \( (X + c)_{\lambda_e} \) are best approximated by normal distributions with equal variances in the family of transformed distributions. Let \( \hat{\lambda} \) be the estimate of \( \lambda_e \) based on the observed samples \( X_1, X_2, \ldots, X_{n_1} \) and \( Y_1, Y_2, \ldots, Y_{n_2} \). Then \( \hat{\lambda} \) maximizes the likelihood function

\[
L_e(\lambda) = (2\pi)^{-n_1/2}\sigma_{\lambda}^{-n_1}e^{-\frac{\sum(X_{\lambda,i} - \mu_{\lambda,i})^2}{2\sigma_{\lambda}^2}} \\
\times (2\pi)^{-n_2/2}\sigma_{\lambda}^{-n_2}e^{-\frac{\sum(Y_{\lambda,i} - \mu_{\lambda,i})^2}{2\sigma_{\lambda}^2}} \times (\Pi_i X_i)^{\lambda-1}(\Pi_i Y_i)^{\lambda-1},
\]
where $\mu_{X_\lambda}$ and $\mu_{Y_\lambda}$ are the means and $\sigma^2_\lambda$ is the common variance in the normal approximation. Let $\hat{\sigma}^2_\lambda = n^{-1}[\sum_{i=1}^{n_1} (X_{\lambda,i} - \bar{X}_\lambda)^2 + \sum_{i=1}^{n_2} (Y_{\lambda,i} - \bar{Y}_\lambda)^2]$. Then $\hat{\lambda}_e$ is the minimizer of

$$l_e(\lambda) = \hat{\sigma}^2_\lambda \exp\{-2n^{-1}\lambda[\sum \log(X_i) + \sum \log(Y_i)]\}.$$ 

Suppose $\lambda_{e0}$ is the unique minimizer of the function

$$l_{e0}(\lambda) = (k + 1)^{-1}[k Var X_\lambda + Var(X + c)_\lambda]$$

$$\times \exp\{-2\lambda[kE \log(X) + E \log(X + c)]/(k + 1)\}$$

over $[a, b]$. Then a proof similar to that of Lemma 1 in Chen and Loh (1992) gives

**Lemma 2** Under conditions A1, A2 and A3, and as $n_1$ and $n_2 \to \infty$,

(i) $l_e(\lambda) \to l_{e0}(\lambda)$ a.s. uniformly in $\lambda$ over $[a, b]$,

(ii) $\hat{\lambda}_e \to \lambda_{e0}$ a.s.

### 2.2 Normal approximation with unequal variances

Since the distributions of $Y$ and $X$ differ only by a location shift, the transformed distributions of $X_\lambda$ and $Y_\lambda$ will not in general have equal variances. Thus another way is to approximate the distributions of $X_\lambda$ and $Y_\lambda$ by normal distributions with unequal variances. The implication of this transformation is similar to the case of normal approximation with equal variances, i.e., there is a $\lambda_u$ such that the distributions of $X_{\lambda_u}$ and $(X + c)_{\lambda_u}$ are best approximated by normal distributions with unequal variances. Let $\hat{\lambda}_u$ be the estimate of $\lambda_u$ based on the observed samples $X_1, X_2, \ldots, X_{n_1}$ and $Y_1, Y_2, \ldots, Y_{n_2}$. Then $\hat{\lambda}_u$ maximizes the likelihood function

$$L_u(\lambda) = (2\pi)^{-n_1/2} \sigma_{X_\lambda}^{-n_1} e^{-\sum(X_{\lambda,i} - \mu_{X_\lambda})^2/2\sigma_{X_\lambda}^2}$$
\[ \times (2\pi)^{-n_2/2} \sigma_{\lambda}^{-n_2} e^{-\frac{\sum(Y_{\lambda,i} - \mu_{\lambda})^2}{2\sigma_{\lambda}^2}} \times (\Pi_i \lambda^1) \lambda^{-1} (\Pi_i Y_i) \lambda^{-1}, \]

where \( \mu_{X_{\lambda}} \) and \( \mu_{Y_{\lambda}} \) are the means, \( \sigma_{X_{\lambda}}^2 \) and \( \sigma_{Y_{\lambda}}^2 \) are the variances in the normal approximation.

Let \( \hat{\sigma}_{X_{\lambda}}^2 = n_1^{-1} \sum_{i=1}^{n_1} (X_{\lambda,i} - \bar{X}_{\lambda})^2 \) and \( \hat{\sigma}_{Y_{\lambda}}^2 = n_2^{-1} \sum_{i=1}^{n_2} (Y_{\lambda,i} - \bar{Y}_{\lambda})^2 \). Then \( \lambda_u \) is the minimizer of

\[ l_u(\lambda) = (\hat{\sigma}_{X_{\lambda}}^2)^{n_1/n} (\hat{\sigma}_{Y_{\lambda}}^2)^{n_2/n} \exp\{ -2n^{-1} \lambda [\sum \log(X_i) + \sum \log(Y_i)] \}. \]

Suppose that \( \lambda_{u0} \) is the unique minimizer of the function

\[ l_{u0}(\lambda) = [(Var.X_{\lambda})^k Var(X+c)\lambda \exp\{ -2\lambda [k E \log(X) + E \log(X+c)] \}]^{1/(k+1)}, \]

over \([a,b]\).

**Lemma 3** Under conditions A1, A2 and A3, and as \( n_1 \) and \( n_2 \to \infty \),

(i) \( l_u(\lambda) \to l_{u0}(\lambda) \) a.s. uniformly in \( \lambda \) over \([a,b]\),

(ii) \( \lambda_u \to \lambda_{u0} \) a.s.

3 Discriminant analysis on transformed data

Upon obtaining the estimated transformation parameter, we can apply LDR or QDR on the transformed data according to the method of estimating the parameter \( \lambda \). For transformations using normal approximation with equal variances, LDR will be applied to the transformed data, and for normal approximation with unequal variances, QDR will be used.

Let \( \lambda_0 \) be the limit of the estimate \( \hat{\lambda} \) in the normal approximation with equal or unequal variances. In order to derive certain properties, suppose that \( \lambda_0 \) is an internal point of the interval \([a,b]\) defined in A1. Without loss of generality, assume that
A4. Let $\delta > 0$ be such that $(\lambda_0 - \delta, \lambda_0 + \delta) \subset (a, b)$.

Lemma 4 Under conditions A1–A4, and as $n_1$ and $n_2 \to \infty$,

(i) $\tilde{X}_\lambda \to EX_{\lambda_0}$ and $\tilde{\sigma}_{\tilde{X}_\lambda}^2 \to Var X_{\lambda_0}$ a.s.,

(ii) $\tilde{Y}_\lambda \to E(X + c)_{\lambda_0}$ and $\tilde{\sigma}_{\tilde{Y}_\lambda}^2 \to Var (X + c)_{\lambda_0}$ a.s.

Proof. We will only prove that $n_1^{-1} \sum X_{\lambda_0,i}^2 - n_1^{-1} \sum X_{\lambda_0,i}^2 \to 0$ a.s. since the proof of the rest of the lemma is similar. Since $X_{\lambda}$ is increasing in $\lambda$, for arbitrary positive $\varepsilon < \delta$, and for $n \geq N$ with $|\tilde{\lambda} - \lambda_0| \leq \varepsilon$,

$$|n_1^{-1} \sum X_{\lambda,i}^2 - n_1^{-1} \sum X_{\lambda_0,i}^2|$$

$$\leq n_1^{-1} \sum |X_{\lambda,i} - X_{\lambda_0,i}|^2$$

$$\leq 2n_1^{-1} \sum (|X_{\lambda_0+\varepsilon,i} - X_{\lambda_0-\varepsilon,i}| + |X_{\lambda_0-\varepsilon,i} - X_{\lambda_0+\varepsilon,i}|)$$

$$\to 2E(|X_{\lambda_0+\varepsilon} - X_{\lambda_0-\varepsilon}|)$$

An application of the Lebesgue dominated convergence theorem and the Strong Law of Large Numbers finishes the proof. \qed

3.1 Equal variances

If the data are transformed using normal approximations with equal variances, then LDR will be used. The LDR on the transformed data is based on $W_{\tilde{\lambda}_e} = Z_{\tilde{\lambda}_e} - (\tilde{X}_{\tilde{\lambda}_e} + \tilde{Y}_{\tilde{\lambda}_e})/2$. Since $X_{\lambda}$ is increasing in $X$, the LDR on the transformed data assigns $Z$ to population 2 when $W_{\tilde{\lambda}_e} = Z_{\tilde{\lambda}_e} - (\tilde{X}_{\tilde{\lambda}_e} + \tilde{Y}_{\tilde{\lambda}_e})/2$ is non-negative and to population 1 otherwise. The corresponding error probabilities are $P[W_{\tilde{\lambda}_e} \geq 0|Z \sim F(x)]$ and $P[W_{\tilde{\lambda}_e} < 0|Z \sim F(x - c)]$. Let $\tilde{p}_e$ be the sum of these
two error probabilities. Define

\[ b = \begin{cases} 
\frac{[EX^{\lambda_0} + E(X + c)^{\lambda_0}]/2}{\lambda_0}, & \text{if } \lambda_0 \neq 0, \\
\exp\{E\log(X) + E\log(X + c)/2\}, & \text{otherwise},
\end{cases} \]

and

\[ p_e = P[|X - (b - c/2)| > c/2]. \tag{2} \]

Then by Slutsky’s Theorem, we have

**Theorem 1** Under conditions A1–A4, \( \hat{p}_e \to p_e \) a.s.

### 3.2 Unequal variances

If normal approximation with unequal variances is used on the transformed data, then QDR will be used. Before looking into the analysis, let us recall the QDR for two normal populations \( N(\mu_1, \sigma_1^2) \) and \( N(\mu_2, \sigma_2^2) \) with density functions \( f_1(x; \mu_1, \sigma_1^2) \) and \( f_2(x; \mu_2, \sigma_2^2) \) respectively, where \( \sigma_1^2 \neq \sigma_2^2 \).

For a given observation \( x \), QDR is based on the likelihood ratio

\[ L(x) = \frac{f_2(x; \mu_2, \sigma_2^2)}{f_1(x; \mu_1, \sigma_1^2)} = \frac{(\sigma_1/\sigma_2)}{2\sigma_1^2} \exp \left\{ \frac{(x - \mu_1)^2}{2\sigma_1^2} - \frac{(x - \mu_2)^2}{2\sigma_2^2} \right\}, \]

which assigns \( x \) to population \( N(\mu_2, \sigma_2^2) \) if \( L(x) \geq 1 \) and to population \( N(\mu_1, \sigma_1^2) \) otherwise. The logarithm of the likelihood ratio is

\[ \frac{\sigma_2^2 - \sigma_1^2}{2\sigma_1^2\sigma_2^2} \left[ x + \frac{\sigma_2^2 \mu_1 - \sigma_1^2 \mu_2}{\sigma_1^2 - \sigma_2^2} \right]^2 + \frac{(\mu_1 - \mu_2)^2}{2(\sigma_1^2 - \sigma_2^2)} - \frac{1}{2} \log(\frac{\sigma_2^2}{\sigma_1^2}). \]
So the inequality $L(x) \geq 1$ is equivalent to

$$\left[ x + \frac{\sigma_2^2 \mu_1 - \sigma_1^2 \mu_2}{\sigma_1^2 - \sigma_2^2} \right]^2 \leq \frac{2\sigma_1^2 \sigma_2^2}{\sigma_1^2 - \sigma_2^2} \left[ \frac{(\mu_1 - \mu_2)^2}{2(\sigma_1^2 - \sigma_2^2)} - \frac{1}{2} \log \left( \frac{\sigma_2^2}{\sigma_1^2} \right) \right], \quad \text{if } \sigma_1^2 > \sigma_2^2$$

and

$$\left[ x + \frac{\sigma_2^2 \mu_1 - \sigma_1^2 \mu_2}{\sigma_1^2 - \sigma_2^2} \right]^2 \geq \frac{2\sigma_1^2 \sigma_2^2}{\sigma_1^2 - \sigma_2^2} \left[ \frac{(\mu_1 - \mu_2)^2}{2(\sigma_1^2 - \sigma_2^2)} - \frac{1}{2} \log \left( \frac{\sigma_2^2}{\sigma_1^2} \right) \right], \quad \text{if } \sigma_1^2 < \sigma_2^2.$$

If the parameters in the two normal distributions are unknown, then one will use estimates of those parameters in the above rule.

Note that if $\sigma_1^2 = \sigma_2^2$, then the MLR (maximum likelihood rule) will reduce to the LDR. If the estimates of $\sigma_1^2$ and $\sigma_2^2$ are equal, then one will use LDR instead of QDR since both rules are MLR under different situations.

In the following, we assume that $\text{Var}X_{\lambda_0} \neq \text{Var}Y_{\lambda_0}$ in the consideration of QDR. Thus with probability one, if $n$ is large enough, the estimated sample variances $\hat{\sigma}_x^2$ and $\hat{\sigma}_y^2$ of the transformed data will be unequal also. QDR will assign $Z$ to population 2 if

$$Q_{\lambda_u} = \frac{\hat{\sigma}_y^2 - \hat{\sigma}_x^2}{2\hat{\sigma}_x^2 \hat{\sigma}_y^2} \left[ Z_{\lambda_u} + \frac{\hat{\sigma}_y^2 \bar{X}_{\lambda_u} - \hat{\sigma}_x^2 \bar{Y}_{\lambda_u}}{\hat{\sigma}_x^2 - \hat{\sigma}_y^2} \right]^2 + \frac{(\bar{X}_{\lambda_u} - \bar{Y}_{\lambda_u})^2}{2(\hat{\sigma}_x^2 - \hat{\sigma}_y^2)} - \frac{1}{2} \log \left( \frac{\hat{\sigma}_y^2}{\hat{\sigma}_x^2} \right)$$

is nonnegative and to population 1 otherwise. Let $\hat{p}_u$ be the sum of error probabilities using $Q_{\lambda_u}$, i.e.,

$$\hat{p}_u = P[Q_{\lambda_u} \geq 0|Z \sim F(x)] + P[Q_{\lambda_u} < 0|Z \sim F(x - c)].$$

Let

$$Q_0 = \frac{(EX_{\lambda_0} - EY_{\lambda_0})^2}{2(\text{Var}X_{\lambda_0} - \text{Var}Y_{\lambda_0})} - \frac{1}{2} \log \left( \frac{\text{Var}Y_{\lambda_0}}{\text{Var}X_{\lambda_0}} \right)$$

$$+ \frac{\text{Var}Y_{\lambda_0} - \text{Var}X_{\lambda_0}}{2\text{Var}X_{\lambda_0} \text{Var}Y_{\lambda_0}} \left[ Z_{\lambda_0} + \frac{\text{Var}Y_{\lambda_0} EX_{\lambda_0} - \text{Var}X_{\lambda_0} EY_{\lambda_0}}{\text{Var}X_{\lambda_0} - \text{Var}Y_{\lambda_0}} \right]^2.$$
By Lemma 2, \( Q_{\lambda_0} \to Q_0 \) a.s. under conditions A1–A4.

Note that \( Q_0 \) is the QDR when the transformed distributions of \( X \) and \( Y \) using \( \lambda_0 \) are approximated by \( N(EX_{\lambda_0}, VarX_{\lambda_0}) \) and \( N(EX_{\lambda_0}, VarX_{\lambda_0}) \). The two error probabilities using \( Q_0 \) are \( P[Q_0 \geq 0|Z \sim F(x)] \) and \( P[Q_0 < 0|Z \sim F(x - c)] \). Denote

\[
\begin{align*}
    a_1 &= (VarX_{\lambda_0}EX_{\lambda_0} - VarY_{\lambda_0}EX_{\lambda_0})/(VarX_{\lambda_0} - VarY_{\lambda_0}) \\
    a_2 &= (2VarX_{\lambda_0}VarY_{\lambda_0})/(VarX_{\lambda_0} - VarY_{\lambda_0}) \\
    a_3 &= (EX_{\lambda_0} - EX_{\lambda_0})^2/[2(VarX_{\lambda_0} - VarY_{\lambda_0})] - (1/2)\log(VarY_{\lambda_0}/VarX_{\lambda_0}).
\end{align*}
\]

If \( VarX_{\lambda_0} > VarY_{\lambda_0} \), then \( Q_0 \geq 0 \) is equivalent to

\[
a_1 - (a_2a_3)^{1/2} \leq Z_{\lambda_0} \leq a_1 + (a_2a_3)^{1/2},
\quad (3)
\]

and if \( VarX_{\lambda_0} < VarY_{\lambda_0} \), then \( Q_0 \leq 0 \) is equivalent to (3). In order to analyze the error probability of QDR, we need to look at the comparison of \( VarX_{\lambda_0} \) and \( VarY_{\lambda_0} \). First, we need the following assumption:

A5. (i) If \( \lambda_0 \neq 0 \), assume that \( EX^{\lambda_0-1} \) and \( EX^{2\lambda_0-1} \) are finite.

(ii) If \( \lambda_0 = 0 \), assume that \( E[\log(X)]^2, EX^{-1} \) and \( EX^{-1}\log(X) \) are finite.

As with conditions A1, A2 and A3, the validity of condition A5 implies the same with \( X \) replaced by \( X + c \) for any \( c > 0 \).

**Lemma 5** Under conditions A1 and A5, \( VarX_{\lambda_0} \geq VarY_{\lambda_0} \) if \( \lambda_0 \leq 1 \), and \( VarX_{\lambda_0} \leq VarY_{\lambda_0} \) otherwise.

**Proof.** Note that for \( \lambda_0 \neq 0 \), \( VarX_{\lambda_0} = VarX^{\lambda_0}/\lambda_0^2 \), and for \( \lambda_0 = 0 \), \( VarX_{\lambda_0} = Var\log(X) \).

(i) Suppose that \( \lambda_0 \leq 1 \) and \( \lambda_0 \neq 0 \). Then \( VarX_{\lambda_0} \geq VarY_{\lambda_0} \) if and only if \( VarX^{\lambda_0} \geq \)
\[\text{Var}Y^{\lambda_0} \text{ which is equivalent to } f(c) \geq 0, \text{ where } f(c) = \text{Var}X^{\lambda_0} - \text{Var}Y^{\lambda_0}. \text{ Note that } f(0) = 0 \text{ and } f'(c) = -\partial E[(X + c)^{2\lambda_0}/\partial c + \partial[E(X + c)^{\lambda_0}]^2/\partial c]. \text{ By the Mean Value Theorem, } \partial E(X + c)^{\lambda_0}/\partial c = \lim_{t \to c} \lambda_0 E(X + c_*^{\lambda_0 - 1}, \text{ where } c_* \text{ is between } t \text{ and } c. \text{ Let } c_0 \text{ be such that } 0 < c_0 < c. \text{ Then when } t \text{ is sufficiently close to } c, \text{ } t \text{ will be greater than } c_0. \text{ So } (X + c_*)^{\lambda_0 - 1} \leq (X + c_0)^{\lambda_0 - 1}. \text{ By assumption (i) in A5, } (X + c_0)^{\lambda_0 - 1} \text{ is integrable. So by the Lebesgue dominated convergence theorem, the order of limit and integration can be changed. Thus}

\[
\partial E(X + c)^{\lambda_0}/\partial c = \lambda_0 E[X + c_*^{\lambda_0 - 1} = \lambda_0 E(X + c)^{\lambda_0 - 1} = E\partial(X + c)^{\lambda_0}/\partial c.
\]

Similarly, the order of differentiation with respect to } c \text{ and expectation for } (X + c)^{2\lambda_0} \text{ can be changed. Therefore}

\[
f'(c) = -[E2\lambda_0(X + c)^{2\lambda_0 - 1} - 2E(X + c)^{\lambda_0}E\lambda_0(X + c)^{\lambda_0 - 1}]
\]

\[= -2\lambda_0 \text{Cov}[(X + c)^{\lambda_0}, (X + c)^{\lambda_0 - 1}].\]

For } 0 < \lambda_0 \leq 1, \text{ } (x + c)^{\lambda_0} \text{ is an increasing function of } x, \text{ and } (x + c)^{\lambda_0 - 1} \text{ is a decreasing function of } x. \text{ Therefore Cov}[(X + c)^{\lambda_0}, (X + c)^{\lambda_0 - 1}] \leq 0. \text{ Thus } f'(c) \geq 0.

For } \lambda_0 < 0, \text{ } (x + c)^{\lambda_0} \text{ and } (x + c)^{\lambda_0 - 1} \text{ are both decreasing functions of } x. \text{ So } f'(c) \geq 0 \text{ also. Hence } f'(c) \geq 0 \text{ for } c \geq 0. \text{ Thus } f(c) \geq 0 \text{ for } c \geq 0 \text{ since } f(0) = 0.

(ii) If } \lambda_0 = 0, \text{ then } X_{\lambda_0} = \log(X) \text{ and } (X + c)_{\lambda_0} = \log(X + c). \text{ The condition } \text{Var}X_{\lambda_0} \geq \text{Var}Y_{\lambda_0} \text{ is equivalent to } f(c) \geq 0, \text{ where } f(c) = \text{Var}\log(X) - \text{Var}\log(X + c). \text{ Note that } f'(c) = -\partial E[\log(X + c)]^2/\partial c + \partial[\log(X + c)]^2/\partial c. \text{ Again by the Mean Value Theorem, } \partial E[\log(X + c)]^2/\partial c = \lim_{t \to c} 2E[\log(X + c_*)]/(X + c_*), \text{ where } c_* \text{ is between } t \text{ and } c. \text{ Let } c_0 \text{ be as before. Then}

\[||\log(X + c_*))/(X + c_*|| \leq ||\log(X + c_0)|| + ||\log(X + c)||/X\]

13
if $t$ is sufficiently close to $c$. Again by (ii) of condition A5 and the Lebesgue dominated convergence theorem, we see that $f'(c) \geq 0$ for $c \geq 0$. Thus $f(c) \geq 0$ for $c \geq 0$ since $f(0) = 0$.

A similar argument holds for the case $\lambda_0 > 1$.

Consider the case $\lambda_0 \leq 1$. Let $d_1 = a_1 - (a_2a_3)^{1/2}$ and $d_2 = a_1 + (a_2a_3)^{1/2}$. Then $Q_0 \geq 0$ if and only if $d_1 \leq Z_{\lambda_0} \leq d_2$. Let $g_\lambda(y) = (\lambda y + 1)^{1/\lambda}$ for $\lambda \neq 0$, and $g_\lambda(y) = \exp(y)$ for $\lambda = 0$, be the inverse function of $y = x_\lambda$. Let $u_1 = g_{\lambda_0}(d_1)$ and $u_2 = g_{\lambda_0}(d_2)$. Then $Q_0 \geq 0$ if and only if $u_1 \leq Z \leq u_2$. Define

$$
\begin{align*}
q_u &= P[u_1 < Z < u_2 | Z \sim F(x)] \\
&= P[Z < u_1 | Z \sim F(x - c)] + P[Z > u_2 | Z \sim F(x - c)] \\
&= P(u_1 < X < u_2) + P(X < u_1 - c) + P(X > u_2 - c).
\end{align*}
$$

(4)

For $\lambda_0 > 1$, $Q_0 \geq 0$ if and only if $Z_{\lambda_0} \leq d_1$ or $Z_{\lambda_0} \geq d_2$. Define

$$
q'_u = P(X < u_1) + P(X > u_2) + P(u_1 - c < X < u_2 - c).
$$

(5)

**Theorem 2** Under conditions A1–A5, $q_u \to p_u$ a.s. for $\lambda_0 \leq 1$, and $q_u \to p'_u$ a.s. for $\lambda_0 > 1$.

4 Numerical comparison of limiting error probabilities

Examples of limiting error probabilities for several distributions will be presented here. The limiting error probabilities before and after transformation are (1), (2), (4) and (5) respectively. Since the limit of $\hat{\lambda}$ depends on the limit $k = \lim n_1/n_2$, the limiting error probabilities depend on $k$ also. In the following comparison of $p$, $p_o$ and $p_u$, the values $k = 1/2, 1,$ and $2$ will be used.
1. Lognormal distribution. Table 1 gives values for the lognormal($\mu, 1$) distribution. For any $\lambda \in [a, b]$, one can apply Box-Cox transformations and calculate the resulting limiting error probability of LDR and QDR on the transformed data. As a function of $\lambda$, these limiting error probabilities are plotted vs $\lambda$ in Figure 1 along with the limiting error probabilities $p$, $p_c$ and $p_u$ for $k = 1$ and several choices of $c$ and ($\mu, 1$). ($\lambda_0 = 1$ corresponds to the untransformed case.)

2. Gamma distribution. Table 2 is the counterpart of table 1 for the gamma($\alpha, 1$) distribution including the exponential distribution ($\alpha = 1$). Figure 2 is the counterpart of Figure 1. ($\lambda_0 = 1$ corresponds to the untransformed case.)

3. Log-double exponential distribution. Table 3 gives values for the log-double exponential($\alpha, \beta$) distribution with $0 < \beta < 1$ ($X$ is log-double exponential if the density function of log $X$ is $p(x) = \exp\{-|x - \alpha|/\beta\}/2\beta$, where $\beta > 0$). Figure 3 is the counterpart of Figure 1. ($\lambda_0 = 1$ corresponds to the untransformed case.)

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$c$</th>
<th>$p$</th>
<th>$k = 0.5$</th>
<th>$k = 1$</th>
<th>$k = 2$</th>
<th>$k = 0.5$</th>
<th>$k = 1$</th>
<th>$k = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$p_u$</td>
<td>$\lambda_u$</td>
<td>$p_u$</td>
<td>$\lambda_u$</td>
<td>$p_u$</td>
<td>$\lambda_u$</td>
</tr>
<tr>
<td>0</td>
<td>0.3</td>
<td>.936</td>
<td>.826</td>
<td>-.15</td>
<td>.829</td>
<td>-.10</td>
<td>.830</td>
<td>-.06</td>
</tr>
<tr>
<td>0</td>
<td>0.8</td>
<td>.825</td>
<td>.567</td>
<td>-.24</td>
<td>.572</td>
<td>-.15</td>
<td>.577</td>
<td>-.08</td>
</tr>
<tr>
<td>0</td>
<td>1.2</td>
<td>.728</td>
<td>.420</td>
<td>-.26</td>
<td>.418</td>
<td>-.16</td>
<td>.418</td>
<td>-.09</td>
</tr>
<tr>
<td>0</td>
<td>1.8</td>
<td>.561</td>
<td>.318</td>
<td>-.28</td>
<td>.313</td>
<td>-.16</td>
<td>.310</td>
<td>-.09</td>
</tr>
<tr>
<td>0</td>
<td>2.5</td>
<td>.323</td>
<td>.238</td>
<td>-.19</td>
<td>.236</td>
<td>-.17</td>
<td>.234</td>
<td>-.09</td>
</tr>
<tr>
<td>1</td>
<td>1.0</td>
<td>.921</td>
<td>.788</td>
<td>-.07</td>
<td>.791</td>
<td>-.11</td>
<td>.796</td>
<td>-.07</td>
</tr>
<tr>
<td>1</td>
<td>2.0</td>
<td>.839</td>
<td>.597</td>
<td>-.23</td>
<td>.603</td>
<td>-.14</td>
<td>.607</td>
<td>-.08</td>
</tr>
<tr>
<td>1</td>
<td>3.0</td>
<td>.752</td>
<td>.448</td>
<td>-.26</td>
<td>.449</td>
<td>-.16</td>
<td>.450</td>
<td>-.09</td>
</tr>
<tr>
<td>1</td>
<td>4.0</td>
<td>.656</td>
<td>.367</td>
<td>-.27</td>
<td>.361</td>
<td>-.16</td>
<td>.356</td>
<td>-.09</td>
</tr>
<tr>
<td>1</td>
<td>5.0</td>
<td>.548</td>
<td>.313</td>
<td>-.28</td>
<td>.308</td>
<td>-.16</td>
<td>.304</td>
<td>-.09</td>
</tr>
</tbody>
</table>

Table 1: Limiting error probabilities for the lognormal($\mu, 1$) distribution.
Table 2: Limiting error probabilities for the gamma(α, 1) distribution including the exponential distribution (α = 1).

<table>
<thead>
<tr>
<th>α</th>
<th>c</th>
<th>p</th>
<th>( k = 0.5 )</th>
<th>( k = 1 )</th>
<th>( k = 2 )</th>
<th>( k = 0.5 )</th>
<th>( k = 1 )</th>
<th>( k = 2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>( p_u )</td>
<td>( \lambda_u )</td>
<td>( p_u )</td>
<td>( \lambda_u )</td>
<td>( p_e )</td>
<td>( \lambda_e )</td>
</tr>
<tr>
<td>1</td>
<td>0.3</td>
<td>.889</td>
<td>.801</td>
<td>.14</td>
<td>.805</td>
<td>.19</td>
<td>.807</td>
<td>.22</td>
</tr>
<tr>
<td>1</td>
<td>0.8</td>
<td>.698</td>
<td>.492</td>
<td>.12</td>
<td>.502</td>
<td>.18</td>
<td>.509</td>
<td>.22</td>
</tr>
<tr>
<td>1</td>
<td>1.2</td>
<td>.532</td>
<td>.340</td>
<td>.12</td>
<td>.332</td>
<td>.18</td>
<td>.326</td>
<td>.22</td>
</tr>
<tr>
<td>1</td>
<td>1.8</td>
<td>.245</td>
<td>.228</td>
<td>.12</td>
<td>.223</td>
<td>.19</td>
<td>.220</td>
<td>.22</td>
</tr>
<tr>
<td>3</td>
<td>1.0</td>
<td>.777</td>
<td>.738</td>
<td>.17</td>
<td>.738</td>
<td>.22</td>
<td>.738</td>
<td>.26</td>
</tr>
<tr>
<td>3</td>
<td>2.0</td>
<td>.561</td>
<td>.504</td>
<td>.13</td>
<td>.505</td>
<td>.20</td>
<td>.505</td>
<td>.25</td>
</tr>
<tr>
<td>3</td>
<td>3.0</td>
<td>.365</td>
<td>.319</td>
<td>.11</td>
<td>.319</td>
<td>.20</td>
<td>.319</td>
<td>.25</td>
</tr>
<tr>
<td>3</td>
<td>4.0</td>
<td>.205</td>
<td>.187</td>
<td>.11</td>
<td>.187</td>
<td>.20</td>
<td>.187</td>
<td>.25</td>
</tr>
</tbody>
</table>

Table 3: Limiting error probabilities for the log-double exponential(\(\alpha, \beta\)) distribution with \(0 < \beta < 1\).

<table>
<thead>
<tr>
<th>(\alpha)</th>
<th>(\beta)</th>
<th>c</th>
<th>p</th>
<th>( k = 0.5 )</th>
<th>( k = 1 )</th>
<th>( k = 2 )</th>
<th>( k = 0.5 )</th>
<th>( k = 1 )</th>
<th>( k = 2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>( p_u )</td>
<td>( \lambda_u )</td>
<td>( p_u )</td>
<td>( \lambda_u )</td>
<td>( p_e )</td>
<td>( \lambda_e )</td>
</tr>
<tr>
<td>0</td>
<td>.3</td>
<td>.5</td>
<td>.474</td>
<td>.438</td>
<td>.13</td>
<td>.440</td>
<td>.16</td>
<td>.435</td>
<td>.06</td>
</tr>
<tr>
<td>0</td>
<td>.3</td>
<td>.8</td>
<td>.281</td>
<td>.258</td>
<td>.18</td>
<td>.258</td>
<td>.19</td>
<td>.256</td>
<td>.07</td>
</tr>
<tr>
<td>0</td>
<td>.3</td>
<td>1.0</td>
<td>.195</td>
<td>.185</td>
<td>.21</td>
<td>.185</td>
<td>.20</td>
<td>.183</td>
<td>.08</td>
</tr>
<tr>
<td>0</td>
<td>.3</td>
<td>1.2</td>
<td>.162</td>
<td>.124</td>
<td>.23</td>
<td>.134</td>
<td>.20</td>
<td>.133</td>
<td>.09</td>
</tr>
<tr>
<td>1</td>
<td>.3</td>
<td>.5</td>
<td>.791</td>
<td>.791</td>
<td>.13</td>
<td>.787</td>
<td>.19</td>
<td>.784</td>
<td>.06</td>
</tr>
<tr>
<td>1</td>
<td>.3</td>
<td>1.5</td>
<td>.434</td>
<td>.405</td>
<td>.26</td>
<td>.400</td>
<td>.16</td>
<td>.398</td>
<td>.10</td>
</tr>
<tr>
<td>1</td>
<td>.3</td>
<td>2.0</td>
<td>.315</td>
<td>.291</td>
<td>.29</td>
<td>.289</td>
<td>.18</td>
<td>.287</td>
<td>.10</td>
</tr>
<tr>
<td>1</td>
<td>.3</td>
<td>3.0</td>
<td>.161</td>
<td>.158</td>
<td>.33</td>
<td>.156</td>
<td>.20</td>
<td>.155</td>
<td>.11</td>
</tr>
<tr>
<td>1</td>
<td>.5</td>
<td>1.0</td>
<td>.838</td>
<td>.746</td>
<td>.19</td>
<td>.738</td>
<td>.12</td>
<td>.733</td>
<td>.07</td>
</tr>
<tr>
<td>1</td>
<td>.5</td>
<td>3.0</td>
<td>.446</td>
<td>.313</td>
<td>.27</td>
<td>.310</td>
<td>.17</td>
<td>.308</td>
<td>.10</td>
</tr>
<tr>
<td>1</td>
<td>.5</td>
<td>5.0</td>
<td>.184</td>
<td>.156</td>
<td>.30</td>
<td>.153</td>
<td>.18</td>
<td>.151</td>
<td>.10</td>
</tr>
<tr>
<td>1</td>
<td>.5</td>
<td>7.0</td>
<td>.074</td>
<td>.100</td>
<td>.30</td>
<td>.099</td>
<td>.18</td>
<td>.098</td>
<td>.10</td>
</tr>
</tbody>
</table>
Figure 1: Plots of limiting error probabilities of LDR and QDR vs $\lambda$ for the lognormal distribution with $k = 1$. The solid line refers to LDR and the dotted line to QDR. The value of $p$ for LDR without transformation is denoted by a "+", the value of $p_c$ for LDR on transformed data by an "c" and the value of $p_u$ for QDR on transformed data by a "u".
Figure 2: Plots of limiting error probabilities of LDR and QDR vs $\lambda$ for the gamma distribution with $k = 1$. The solid line refers to LDR and the dotted line for QDR. The value of $p$ for LDR without transformation is denoted by a "+", the value of $p_e$ for LDR on transformed data by an "e" and the value of $p_u$ for QDR on transformed data by a "u".
Figure 3: Plots of limiting error probabilities of LDR and QDR vs $\lambda$ for the log-double exponential distribution with $k = 1$. The solid line refers to LDR and the dotted line for QDR. The value of $p$ for LDR without transformation is denoted by a "+", the value of $p_e$ for LDR on transformed data by an "e" and the value of $p_u$ for QDR on transformed data by a "u".

We see from the results that, for the distributions considered, the use of transformations often produces significant reductions of error probabilities. The QDR for normal approximation with unequal variances is most frequently the best. The LDR without transformations has the largest error probabilities among all the methods.

Another remarkable result about the QDR is that it can reduce error probabilities significantly even when the reduction of error probabilities using LDR on transformed data is not large. In
fact, in most of the plots in Figures 1, 2, and 3, the error probabilities of QDR are very close to the minimum error probabilities using LDR or QDR on transformed data. Of course, the LDR on transformed data also reduces the error probabilities significantly in most cases we consider, especially when location shift $c$ is small. The larger the value of $c$, the easier it is to separate the two populations even without transformation. So the reduction of error probabilities decreases with increasing $c$. Indeed, only when the error probability is large, is there the necessity and possibility of reducing it by other methods.

5 Distributions with unknown left endpoint

It has been assumed so far that the distributions have support only on the positive real line. We now consider the case where the data may take negative values but the support of the underlying distributions have a finite lower bound.

Suppose $F(x)$ has its support $(\theta, \infty)$, where $\theta < 0$ is unknown. Let $\hat{\theta}$ be any strongly consistent estimate of $\theta$ such that $(X_i - \hat{\theta})$'s and $(Y_i - \hat{\theta})$'s are all positive. One possible choice of $\hat{\theta}$ is

$$
\hat{\theta} = \min\{X(1) - [X(2) - X(1)], Y(1) - [Y(2) - Y(1)]\}, \tag{6}
$$

where $X(1) \leq X(2) \leq \ldots \leq X(n_1)$ are the order statistics of $X_1, X_2, \ldots, X_{n_1}$ and $Y(1) \leq Y(2) \leq \ldots \leq Y(n_2)$ are the order statistics of $Y_1, Y_2, \ldots, Y_{n_2}$. The dependence of $X(1)$ and $Y(1)$ on $n$ is suppressed.

The choice of $\hat{\theta}$ in (6) is a version of an estimate proposed by Robson and Whitlock (1964). Intuitively, $X(1)$ is too large as an estimate of $\theta$. The difference between $X(1)$ and $\theta$ can be estimated by that between $X(1)$ and $X(2)$. References for other choices of $\hat{\theta}$ can be found in Loh (1984).
Lemma 6 Suppose that for some \( \varepsilon > 0 \), the \( p \)-quantile \( \zeta_p \) of the distribution \( F(x) \) is unique for any \( p < \varepsilon \). Then \( \hat{\theta} \to \theta \) a.s.

Proof. For any sequence of \( \{X_i\} \), \( X(1) \) is decreasing in \( n \). So there is a random variable \( W \) such that \( X(1) \to W \) a.s. Let \( p < \varepsilon \), and \( \hat{\eta}_p \) be the \( p \)-quantile of the empirical distribution of \( \{X_1, X_2, \ldots, X_{n_1}\} \). Then for large \( n_1 \), \( X(1) \leq \hat{\eta}_p \). Since \( \hat{\eta}_p \to \zeta_p \) a.s., \( \theta \leq W \leq \zeta_p \). Since \( p \) is arbitrary, we have \( W = \theta \) a.s. Similarly \( X(2) \to \theta \) a.s., \( Y(1) \to \theta + c \) a.s. and \( Y(2) \to \theta + c \) a.s. Hence \( \hat{\theta} \to \theta \) a.s. \( \square \)

As in the case \( \theta = 0 \), we will transform the distributions of \( (x - \theta) \lambda \) and \( (y - \theta) \lambda \) to approximate normality for some \( \lambda \). For unequal variances approximation, the estimate \( \hat{\lambda}_{u\theta} \) of \( \lambda \) maximizes

\[
L_{u\theta}(\lambda) = (2\pi)^{-n_1/2} \sigma_{x\lambda}^{-n_1} e^{-\frac{1}{2} \sum [(X_i - \hat{\theta})_\lambda - \mu_{x\lambda}]^2/2\sigma_{x\lambda}^2} \\
\times (2\pi)^{-n_2/2} \sigma_{y\lambda}^{-n_2} e^{-\frac{1}{2} \sum [(Y_i - \hat{\theta})_\lambda - \mu_{y\lambda}]^2/2\sigma_{y\lambda}^2} \times [\Pi_i(X_i - \hat{\theta})]^{\lambda-1} [\Pi_i(Y_i - \hat{\theta})]^{\lambda-1},
\]

where \( \mu_{x\lambda} \) and \( \mu_{y\lambda} \) are the means, \( \sigma_{x\lambda}^2 \) and \( \sigma_{y\lambda}^2 \) are the variances in the normal approximation. Let \( X_{\lambda}(\hat{\theta}) = n_1^{-1} \sum (X_i - \hat{\theta})_\lambda \), \( Y_{\lambda}(\hat{\theta}) = n_2^{-1} \sum (Y_i - \hat{\theta})_\lambda \), \( \sigma_{x\lambda}^2(\hat{\theta}) = n_1^{-1} \sum [(X_i - \hat{\theta})_\lambda - X_{\lambda}(\hat{\theta})]^2 \) and \( \sigma_{y\lambda}^2(\hat{\theta}) = n_2^{-1} \sum [(Y_i - \hat{\theta})_\lambda - Y_{\lambda}(\hat{\theta})]^2 \). Then \( \hat{\lambda}_{u\theta} \) is the minimizer of

\[
l_{u\theta}(\lambda) = [\sigma_{x\lambda}^2(\hat{\theta})]^{n_1/2} [\sigma_{y\lambda}^2(\hat{\theta})]^{n_2/2} \exp\{-2n^{-1}\lambda[\sum \log(X_i - \hat{\theta}) + \sum \log(Y_i - \hat{\theta})]\}).
\]

We need the following conditions instead of previous A1 and A2:

**B1.** There is \( \delta_0 > 0 \) such that \( E(X - \theta)_\lambda^2 < \infty \) for \( \lambda \in (a - \delta_0, b + \delta_0) \),

**B2.** \( E|\log(X - \theta)| < \infty \).

Recall \( l_{u\theta} \) and \( \lambda_{u\theta} \) defined previously with \( \theta = 0 \).

**Theorem 3** Under the assumptions of Lemma 6 and conditions B1, B2 and A3, as \( n_1 \) and \( n_2 \to \)
\( \lambda_\theta \rightarrow \lambda_{\infty} \) a.s.

(i) \( l_\theta(\lambda) \rightarrow l_\theta(\lambda) \) a.s. uniformly in \( \lambda \) over \([a, b]\),

(ii) \( \hat{\lambda}_\theta \rightarrow \lambda_\theta \) a.s.

Obviously, it is enough to prove that uniformly in \( \lambda \in [a, b] \), \( \bar{X}_\lambda(\hat{\theta}) \rightarrow E(X - \theta) \lambda \), \( \bar{Y}_\lambda(\hat{\theta}) \rightarrow E(Y - \theta) \lambda \), \( \sigma^2_{\bar{X}_\lambda}(\hat{\theta}) \rightarrow Var(X - \theta) \lambda \) and \( \sigma^2_{\bar{Y}_\lambda}(\hat{\theta}) \rightarrow Var(Y - \theta) \lambda \). Before we prove the Theorem, we need the following lemma.

**Lemma 7** Let \( U \) be a positive random variable with \( E|U| < \infty \) for some \( \lambda \). Let \( u = \{U_i\} \) be a sequence of i.i.d. random variables distributed as \( U \). Then as \( d \rightarrow \infty \),

\[
\sup_{N \geq 1} N^{-1} \sum_{i=1}^{N} U_{\lambda_i} I(U_i \geq d) \rightarrow 0
\]

for almost every sequence \( u \).

**Proof.** For any \( \epsilon > 0 \), there is \( d_0 > 1 \) such that \( |EU_{\lambda} I(U \geq d_0)| < \epsilon \). Denote \( T_N = \sum_{i=1}^{N} U_{\lambda_i} I(U_i \geq d) \) and \( T_N^0 = \sum_{i=1}^{N} U_{\lambda_i} I(U_i \geq d_0) \). By the Strong Law of Large Numbers, \( T_N^0 \rightarrow EU_{\lambda} I(U \geq d_0) \) a.s. Let \( N_0(u) \) be such that if \( N \geq N_0(u) \), then

\[
EU_{\lambda} I(U \geq d_0) - \epsilon \leq T_N^0 \leq EU_{\lambda} I(U \geq d_0) + \epsilon.
\]

Thus \( -2\epsilon \leq \sup_{N \geq N_0(u)} T_N^0 \leq 2\epsilon \). Since \( I(U \geq d) \) is decreasing in \( d \) and \( U_{\lambda} > 0 \) for \( U > 1 \) by Theorem A.2 part (a) of Hernandez, 1978, \( 0 \leq T_N \leq T_N^0 \leq 2\epsilon \) for \( d \geq d_0 \). Thus for \( d > d_0 \),

\[
-2\epsilon \leq \sup_{N \geq N_0(u)} T_N \leq 2\epsilon.
\]

Let \( d_1(u) = \max_{1 \leq i \leq N_0(u)} \{U_i\} \). Then \( \sup_{N \leq N_0(u)} T_N = 0 \) for \( d > d_1(u) \). Let \( d^*(u) = \max\{d_0, d_1(u)\} \). Then for \( d > d^*(u) \), \( -2\epsilon \leq \sup_{N \geq 1} T_N \leq 2\epsilon \). Hence \( \sup_{N \geq 1} T_N \) converges to 0 as \( d \rightarrow \infty \) almost surely. \( \Box \)
For the sake of simplicity, let \( \omega \) be the random element corresponding to \( \{X_i\} \) and \( \{Y_i\} \). Let \( \hat{\theta} \) be defined in (6), and let \( H_n \) be the empirical distribution of \( \{X_1 - \hat{\theta}, X_2 - \hat{\theta}, \ldots, X_{n_1} - \hat{\theta}\} \). Then by Lemma 2 of Qu and Loh (1992), \( H_n \to F_0 \) in distribution almost surely, where \( F_0 \) is the distribution of \( X - \theta \).

**Proof of Theorem 3.** The proof is quite similar to that of Theorem 2 of Jennrich (1969). For any \( \lambda \in [a, b] \), let \( h_n(\lambda) = n_1^{-1} \sum (X_i - \hat{\theta})_\lambda - E(X - \theta)_\lambda = E_{H_n} X_\lambda - E_{F_0} X_\lambda \). Given \( \delta < \delta_0 \), we have \((\lambda - \delta, \lambda + \delta) \subset (a - \delta_0, b + \delta_0)\) and

\[
\sup_{\hat{\lambda} \in (\lambda - \delta, \lambda + \delta)} h_n(\hat{\lambda}) = \sup_{\hat{\lambda} \in (\lambda - \delta, \lambda + \delta)} \left\{ E_{H_n} X_\hat{\lambda} - E_{F_0} X_\hat{\lambda} \right\} \\
\leq n_1^{-1} \sum_{\hat{\lambda} \in (\lambda - \delta, \lambda + \delta)} \sup_{\hat{\lambda} \in (\lambda - \delta, \lambda + \delta)} (X_i - \hat{\theta})_\hat{\lambda} - E_{F_0} \inf_{\hat{\lambda} \in (\lambda - \delta, \lambda + \delta)} X_\hat{\lambda} \\
= E_{H_n} \sup_{\hat{\lambda} \in (\lambda - \delta, \lambda + \delta)} X_\hat{\lambda} - E_{F_0} \inf_{\hat{\lambda} \in (\lambda - \delta, \lambda + \delta)} X_\hat{\lambda}.
\]

Let \( N_1(\omega) \) be such that \( n \geq N_1(\omega) \) implies \( \theta - \gamma < \hat{\theta} \) for given \( \gamma > 0 \). By the monotonicity of \( X_\lambda \) in \( X \) and Theorem A.2 part (a) of Hernandez, 1978, for \( d > 1 \),

\[
0 \leq |(X_i - \hat{\theta})_{\lambda - \delta} I(X_i - \hat{\theta} \geq d) = (X_i - \hat{\theta})_{\lambda - \delta} I(X_i - \hat{\theta} \geq d) \\
\leq (X_i - \theta + \gamma)_{\lambda - \delta} I(X_i - \theta + \gamma \geq d).
\]

So for \( n \geq N_1(\omega) \) and \( d > 1 \),

\[
E_{H_n} |X_{\lambda - \delta} I(|X| \geq d) = n_1^{-1} \sum |(X_i - \hat{\theta})_{\lambda - \delta} I(|X_i - \hat{\theta}| \geq d) \\
\leq n_1^{-1} \sum (X_i - \theta + \gamma)_{\lambda - \delta} I(X_i - \theta + \gamma \geq d).
\]
Let \( d_1(\omega) = \max_{i \leq N_1(\omega)} \{ X_i - \hat{\theta} \} \). Then for \( d > \max(1, d_1(\omega)) \),

\[
\sup_{n \geq 1} E_{H_n} |X_{\lambda-\delta}| I(|X| \geq d) \leq \sup_{n \geq 1} \sum_{i \geq 1} (X_i - \theta + \gamma)_{\lambda-\delta} I(X_i - \theta + \gamma \geq d).
\]

Since \( E(X - \theta)_{\lambda-\delta} \) is finite, \( E(X - \theta + \gamma)_{\lambda-\delta} \) is finite also by Lemma 1. By Lemma 7, the right hand side of the above inequality goes to 0 as \( d \to \infty \). So \( X_{\lambda-\delta} \) is uniformly integrable relative to \( \{H_n\} \). Similarly, \( X_{\lambda+\delta} \) is uniformly integrable relative to \( \{H_n\} \), and so is \( \sup_{\lambda \in (\lambda-\delta, \lambda+\delta)} X_{\lambda} \). By Lemma 1 of Jennrich, 1969, \( \sup_{\lambda \in (\lambda-\delta, \lambda+\delta)} X_{\lambda} \) is continuous function of \( X \). Therefore by Corollary 8.1.5 of Chow and Teicher (1988), \( \lim_{n \to \infty} E_{H_n} \sup_{\lambda \in (\lambda-\delta, \lambda+\delta)} X_{\lambda} = E_{F_0} \sup_{\lambda \in (\lambda-\delta, \lambda+\delta)} X_{\lambda} \). Hence

\[
\limsup_{n \to \infty} \sup_{\lambda \in (\lambda-\delta, \lambda+\delta)} h_n(\tilde{\lambda}) \leq \int(\sup_{\lambda \in (\lambda-\delta, \lambda+\delta)} X_{\lambda} - \inf_{\lambda \in (\lambda-\delta, \lambda+\delta)} X_{\lambda})dF_0.
\]

Since \( \sup_{\lambda \in (\lambda-\delta, \lambda+\delta)} X_{\lambda} - \inf_{\lambda \in (\lambda-\delta, \lambda+\delta)} X_{\lambda} \to 0 \) as \( \delta \to 0 \),

\[
\lim_{\delta \to 0} \int(\sup_{\lambda \in (\lambda-\delta, \lambda+\delta)} X_{\lambda} - \inf_{\lambda \in (\lambda-\delta, \lambda+\delta)} X_{\lambda})dF_0 = 0
\]

by the Lebesgue dominated convergence theorem because the integrand is dominated by the integrable function \( 2(|X_{\lambda-\delta}| + |X_{\lambda+\delta}|) \). So for given \( \varepsilon > 0 \), there is \( \delta' < \delta \) such that

\[
\limsup_{n \to \infty} \sup_{\lambda \in (\lambda-\delta', \lambda+\delta')} h_n(\tilde{\lambda}) \leq \varepsilon.
\]

Thus there is \( N_\lambda(\omega) \) such that \( n \geq N_\lambda(\omega) \) implies \( \sup_{\lambda \in (\lambda-\delta', \lambda+\delta')} h_n(\tilde{\lambda}) \leq \varepsilon \), which in turn implies \( E_{H_n} X_{\tilde{\lambda}} = n_{\lambda}^{-1} \sum (X_i - \hat{\theta})_{\tilde{\lambda}} \leq E(X - \theta)_{\tilde{\lambda}} + \varepsilon \) for \( n \geq N_\lambda(\omega) \) and \( \tilde{\lambda} \in (\lambda - \delta', \lambda + \delta') \). It follows that \([a, b]\) is covered by such open intervals. By the compactness of \([a, b]\), there is a finite number of such intervals covering \([a, b]\). Let \( N(\omega) \) be the maximum element of the set of integers \( \{N_\lambda(\omega)\} \)
corresponding to these intervals. If \( n \geq N(\omega) \) and \( \lambda \in [a, b] \), we have \( n^{-1} \sum (X_i - \hat{\theta}) \geq E(X - \theta) + \epsilon \). An analogous argument gives \( n^{-1} \sum (X_i - \hat{\theta}) \geq E(X - \theta) - \epsilon \) for large \( n \) and \( \lambda \in [a, b] \).

Thus uniformly in \( \lambda \in [a, b] \), \( n^{-1} \sum (X_i - \hat{\theta}) \to E(X - \theta) \) a.s.

A similar proof holds for \( \bar{Y}_n(\hat{\theta}), \sqrt{\frac{n}{Y_n}}(\hat{\theta}) \) and \( \hat{\theta}^2 \). So \( l_{u\theta}(\lambda) \) approaches \( l_{u\theta}(\lambda) \) uniformly in \( \lambda \in [a, b] \) a.s.

Part (ii) of the theorem is an immediate result of Lemma 2 of Chen, 1990. \( \square \)

Let \( Q_{\hat{\lambda}_{u\theta}} \) be the counterpart of \( Q_{\hat{\lambda}_u} \) with \( X_i \)'s, \( Y_i \)'s and \( \hat{\lambda}_u \) replaced by \( Z - \hat{\theta} \), \( (X_i - \hat{\theta}) \)'s, \( (Y_i - \hat{\theta}) \)'s and \( \hat{\lambda}_{u\theta} \). Then \( Z \) will be assigned to population 2 if \( Z > \hat{\theta} \) and \( Q_{\hat{\lambda}_{u\theta}} \) is nonnegative and to population 1 otherwise. The sum of error probabilities \( \hat{p}_{u\theta} \) using \( Q_{\hat{\lambda}_{u\theta}} \) will approach \( p_u \) or \( p'_u \) according to \( \lambda_{u\theta} \leq 1 \) or not.

A similar result holds if normal approximation with equal variances is used. In this case, one will use the minimizer \( \hat{\lambda}_{e\theta} \) of the counterpart \( l_{e\theta}(\lambda) \) of \( l_e(\lambda) \) with \( X_i \)'s and \( Y_i \)'s replaced by \( (X_i - \hat{\theta}) \)'s and \( (Y_i - \hat{\theta}) \)'s. \( Z \) will be classified to population 2 if \( Z > \hat{\theta} \) and \( W_{\hat{\lambda}_{e\theta}} = (Z - \hat{\theta})_{\hat{\lambda}_{e\theta}} - \bar{X}_{\hat{\lambda}_{e\theta}}(\hat{\theta}) + \bar{Y}_{\hat{\lambda}_{e\theta}}(\hat{\theta}) \) is nonnegative and to population 1 otherwise. The sum of error probabilities \( \hat{p}_{e\theta} \) using \( W_{\hat{\lambda}_{e\theta}} \) will converge to \( p_e \).

**Finite sample simulation**

In order to compare the limiting error probabilities to their finite sample values, a small simulation experiment was conducted for the exponential and lognormal distributions. The two sample sizes \( n_1 \) and \( n_2 \) were chosen equal in the simulation with \( n_1 = n_2 = 30, 50 \) and 100. The simulation was done with both \( \theta = 0 \) and \( \theta < 0 \) but unknown. In the first case, random samples \( X_i \)'s and \( Y_i \)'s were obtained and error probabilities \( \hat{p}_e \) and \( \hat{p}_u \) were calculated using \( W_{\hat{\lambda}_e} \) and \( Q_{\hat{\lambda}_u} \) respectively.

In the second case, samples \( X_i \)'s and \( Y_i \)'s were obtained and \( \theta \) was estimated afterwards, and then error probabilities \( \hat{p}_{e\theta} \) and \( \hat{p}_{u\theta} \) were calculated using \( W_{\hat{\lambda}_{e\theta}} \) and \( Q_{\hat{\lambda}_{u\theta}} \). The number of iterations in
the simulation was 100 for each comparison, and the estimate of various error probabilities were the average of each 100 estimates of error probabilities. The limiting error probabilities \( p, pe \) and \( p_u \) before and after transformations were calculated accurately since the distributions used in simulation were known. The simulation results were presented in table 4 and 5. It is found that accuracy of the error probabilities to their limits was satisfied although the convergence with \( \theta = 0 \) is a bit better than with unknown \( \theta > 0 \). This discrepancy of the performance is of course due to the fact that the unknown \( \theta \) was estimated first before using transformation.

### Table 4: Simulated error probabilities for the standard exponential distribution.

<table>
<thead>
<tr>
<th>( c )</th>
<th>( n_1 = n_2 )</th>
<th>( p )</th>
<th>( pe )</th>
<th>( \hat{pe} )</th>
<th>st.err.</th>
<th>( \hat{pe}_\theta )</th>
<th>st.err.</th>
<th>( p_u )</th>
<th>( \hat{p}_u )</th>
<th>st.err.</th>
<th>( \hat{p}<em>u \hat{pe}</em>\theta )</th>
<th>st.err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>30</td>
<td>.889</td>
<td>.854</td>
<td>.855</td>
<td>.015</td>
<td>.853</td>
<td>.015</td>
<td>.805</td>
<td>.811</td>
<td>.020</td>
<td>.815</td>
<td>.020</td>
</tr>
<tr>
<td>0.3</td>
<td>50</td>
<td>.852</td>
<td>.013</td>
<td>.851</td>
<td>.014</td>
<td>.805</td>
<td>.009</td>
<td>.809</td>
<td>.014</td>
<td>.014</td>
<td>.807</td>
<td>.008</td>
</tr>
<tr>
<td>0.3</td>
<td>100</td>
<td>.853</td>
<td>.009</td>
<td>.853</td>
<td>.008</td>
<td>.805</td>
<td>.009</td>
<td>.807</td>
<td>.008</td>
<td>.008</td>
<td>.807</td>
<td>.008</td>
</tr>
<tr>
<td>0.8</td>
<td>30</td>
<td>.698</td>
<td>.609</td>
<td>.613</td>
<td>.042</td>
<td>.609</td>
<td>.047</td>
<td>.502</td>
<td>.529</td>
<td>.040</td>
<td>.524</td>
<td>.040</td>
</tr>
<tr>
<td>0.8</td>
<td>50</td>
<td>.605</td>
<td>.040</td>
<td>.606</td>
<td>.043</td>
<td>.512</td>
<td>.036</td>
<td>.514</td>
<td>.028</td>
<td>.028</td>
<td>.511</td>
<td>.022</td>
</tr>
<tr>
<td>0.8</td>
<td>100</td>
<td>.606</td>
<td>.025</td>
<td>.608</td>
<td>.025</td>
<td>.505</td>
<td>.023</td>
<td>.511</td>
<td>.022</td>
<td>.022</td>
<td>.505</td>
<td>.022</td>
</tr>
<tr>
<td>1.2</td>
<td>30</td>
<td>.532</td>
<td>.390</td>
<td>.391</td>
<td>.062</td>
<td>.390</td>
<td>.062</td>
<td>.332</td>
<td>.359</td>
<td>.037</td>
<td>.358</td>
<td>.035</td>
</tr>
<tr>
<td>1.2</td>
<td>50</td>
<td>.388</td>
<td>.054</td>
<td>.389</td>
<td>.055</td>
<td>.342</td>
<td>.028</td>
<td>.348</td>
<td>.029</td>
<td>.029</td>
<td>.348</td>
<td>.029</td>
</tr>
<tr>
<td>1.2</td>
<td>100</td>
<td>.390</td>
<td>.036</td>
<td>.384</td>
<td>.035</td>
<td>.336</td>
<td>.016</td>
<td>.335</td>
<td>.018</td>
<td>.018</td>
<td>.335</td>
<td>.018</td>
</tr>
</tbody>
</table>

### Table 5: Simulated error probabilities for the standard lognormal distribution.

<table>
<thead>
<tr>
<th>( c )</th>
<th>( n_1 = n_2 )</th>
<th>( p )</th>
<th>( pe )</th>
<th>( \hat{pe} )</th>
<th>st.err.</th>
<th>( \hat{pe}_\theta )</th>
<th>st.err.</th>
<th>( p_u )</th>
<th>( \hat{p}_u )</th>
<th>st.err.</th>
<th>( \hat{p}<em>u \hat{pe}</em>\theta )</th>
<th>st.err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>30</td>
<td>.825</td>
<td>.691</td>
<td>.692</td>
<td>.043</td>
<td>.673</td>
<td>.049</td>
<td>.572</td>
<td>.546</td>
<td>.110</td>
<td>.592</td>
<td>.142</td>
</tr>
<tr>
<td>0.8</td>
<td>50</td>
<td>.688</td>
<td>.032</td>
<td>.671</td>
<td>.038</td>
<td>.581</td>
<td>.016</td>
<td>.575</td>
<td>.018</td>
<td>.018</td>
<td>.571</td>
<td>.015</td>
</tr>
<tr>
<td>0.8</td>
<td>100</td>
<td>.690</td>
<td>.027</td>
<td>.678</td>
<td>.021</td>
<td>.576</td>
<td>.012</td>
<td>.571</td>
<td>.015</td>
<td>.015</td>
<td>.571</td>
<td>.015</td>
</tr>
<tr>
<td>1.8</td>
<td>30</td>
<td>.561</td>
<td>.257</td>
<td>.315</td>
<td>.055</td>
<td>.308</td>
<td>.045</td>
<td>.313</td>
<td>.348</td>
<td>.040</td>
<td>.384</td>
<td>.049</td>
</tr>
<tr>
<td>1.8</td>
<td>50</td>
<td>.302</td>
<td>.046</td>
<td>.298</td>
<td>.037</td>
<td>.314</td>
<td>.018</td>
<td>.332</td>
<td>.027</td>
<td>.027</td>
<td>.332</td>
<td>.021</td>
</tr>
<tr>
<td>1.8</td>
<td>100</td>
<td>.295</td>
<td>.027</td>
<td>.289</td>
<td>.025</td>
<td>.311</td>
<td>.014</td>
<td>.322</td>
<td>.021</td>
<td>.021</td>
<td>.322</td>
<td>.021</td>
</tr>
<tr>
<td>2.5</td>
<td>30</td>
<td>.323</td>
<td>.211</td>
<td>.213</td>
<td>.025</td>
<td>.225</td>
<td>.029</td>
<td>.236</td>
<td>.234</td>
<td>.020</td>
<td>.252</td>
<td>.030</td>
</tr>
<tr>
<td>2.5</td>
<td>50</td>
<td>.210</td>
<td>.021</td>
<td>.220</td>
<td>.023</td>
<td>.235</td>
<td>.014</td>
<td>.248</td>
<td>.021</td>
<td>.021</td>
<td>.248</td>
<td>.021</td>
</tr>
<tr>
<td>2.5</td>
<td>100</td>
<td>.209</td>
<td>.016</td>
<td>.216</td>
<td>.016</td>
<td>.236</td>
<td>.009</td>
<td>.245</td>
<td>.014</td>
<td>.014</td>
<td>.245</td>
<td>.014</td>
</tr>
</tbody>
</table>
6 Transformations in multivariate situations

In the preceding sections, we studied the application of Box-Cox transformations to univariate populations. It was found that using Box-Cox transformations can reduce error probabilities significantly when parent populations depart from normality. Methodologically, there is no difference between univariate and multivariate populations in applying these transformations. But one difficulty arises with the choice of transformation parameters in the multivariate case. Suppose the dimension of the observed populations is \( m \). The transformation parameters should use all the information contained in each component of the observations. Nevertheless, the common method in practice is for each component of the observations to choose a different \( \lambda \) obtained by considering transforming the marginal distributions of each component to approximate normality. The rationale of this method is that the joint distributions of the transformed distributions will not behave far from normality if each marginal distribution is normal. Because the results in the multivariate case are similar to those in univariate case, we will state them briefly without proof.

Let \( X = (X_1, X_2, \ldots, X_m)' \) and \( Y = (Y_1, Y_2, \ldots, Y_m)' \) be positive observations from populations 1 and 2 respectively, and \( X_1, X_2, \ldots, X_{n_1} \) and \( Y_1, Y_2, \ldots, Y_{n_2} \) be two learning samples from populations 1 and 2 respectively. Applying the normal approximation with equal or unequal variances to the marginal distributions of the \( i \)-th component of \( X \) and \( Y \) yields the estimated transformation parameter \( \hat{\lambda}_i \) for the \( i \)-th components of the samples of \( X_1, X_2, \ldots, X_{n_1} \) and \( Y_1, Y_2, \ldots, Y_{n_2} \). The resulting vector of transformation parameters \( \hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_m)' \) is the one used in applying Box-Cox transformations to normality for the multivariate case. Under similar conditions to A1, A2 and A3, there exists a \( \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)' \) such that \( \lambda \rightarrow \hat{\lambda} \) when \( n_1 \) and \( n_2 \) approach infinity. Let \( X_\lambda = (X_{\lambda,1}, X_{\lambda,2}, \ldots, X_{\lambda,n_1})' \) be the transformation of \( X \) and \( X_{\lambda,1}, X_{\lambda,2}, \ldots, X_{\lambda,n_1} \) the transformations of \( X_1, X_2, \ldots, X_{n_1} \) using \( \lambda \). The sample mean and dispersion matrix of the transformed
The $X$'s are respectively $\bar{X}_\lambda = n_1^{-1} \sum X_{\lambda,i}$ and $\bar{V}_\lambda X = (n_1 - 1)^{-1} \sum (X_{\lambda,i} - \bar{X}_\lambda)(X_{\lambda,i} - \bar{X}_\lambda)'$, where the prime denotes the transpose of a matrix. Similar notations hold for $Y_1, Y_2, \ldots, Y_{n_2}$.

Let $EX_\lambda$, $EY_\lambda$, $\text{Var} X_\lambda$ and $\text{Var} Y_\lambda$ be the means and covariance matrices of the transformed populations. Let $V_\lambda = (k \text{Var} X_\lambda + \text{Var} Y_\lambda)/(k + 1)$, where $k = \lim n_1/n_2$.

If LDR is considered, the pooled estimated covariance matrix is $\bar{V}_\lambda = (n_1 + n_2 - 2)^{-1} \sum (X_{\lambda,i} - \bar{X}_\lambda)(X_{\lambda,i} - \bar{X}_\lambda)' + \sum (Y_{\lambda,i} - \bar{Y}_\lambda)(Y_{\lambda,i} - \bar{Y}_\lambda)'$. The LDR on the transformed data $Z_\lambda$ of a further observation $Z$ assigns $Z$ to population 2 if the discriminant function

$$W_\lambda = [Z_\lambda - .5(\bar{X}_\lambda + \bar{Y}_\lambda)]' \bar{V}_\lambda^{-1} [\bar{Y}_\lambda - \bar{X}_\lambda]$$

is nonnegative and to population 1 otherwise. The error probability of using $W_\lambda$ is

$$\hat{p}_e = P(W_\lambda \geq 0|Z \sim F_1) + P(W_\lambda < 0|Z \sim F_2).$$

If QDR is used, the discriminant function is

$$Q_\lambda = [Z_\lambda - \bar{Y}_\lambda]' \bar{V}_\lambda^{-1} [Z_\lambda - \bar{Y}_\lambda] - [Z_\lambda - \bar{X}_\lambda]' \bar{V}_\lambda^{-1} [Z_\lambda - \bar{X}_\lambda].$$

QDR assigns $Z$ to population 2 or 1 according to whether $Q_\lambda$ is nonnegative or not. The resulting error probability

$$\hat{p}_u = P(Q_\lambda \geq 0|Z \sim F_1) + P(Q_\lambda < 0|Z \sim F_2).$$

Let $W_\lambda$, $p_e$, $Q_\lambda$ and $p_u$ be the counterparts of $W_\lambda$, $\hat{p}_e$, $Q_\lambda$ and $\hat{p}_u$ with $Z_\lambda$, $\bar{X}_\lambda$, $\bar{Y}_\lambda$, $\bar{V}_\lambda X$, $\bar{V}_\lambda Y$, $\bar{V}_\lambda$ and $\lambda$ replaced by $EX$, $EY$, $\text{Var} X_\lambda$, $\text{Var} Y_\lambda$, $V_\lambda$ and $\lambda$ respectively. Assume the following conditions, which are similar to A1–A4.
C1. There are constants $a_i$ and $b_i$ such that $E(X_i^{2,\lambda}) < \infty$ and $E(Y_i^{2,\lambda}) < \infty$ for $\lambda \in [a_i, b_i]$ for $i = 1, 2, \ldots, m$.

C2. $E|\log(X_i)| < \infty$ and $E|\log(Y_i)| < \infty$ for $i = 1, 2, \ldots, m$.

C3. $\lambda_i \in (a_i, b_i)$ and $\hat{\lambda}_i \in [a_i, b_i]$ for $i = 1, 2, \ldots, m$.

**Theorem 4** Under conditions C1–C3, $\hat{p}_e \rightarrow p_e$ and $\hat{p}_u \rightarrow p_u$ a.s. as $n_1$ and $n_2 \rightarrow \infty$. 
References


