MINIMIZING ERROR PROBABILITY OF LINEAR
DISCRIMINANT ANALYSIS USING
BOX-COX TRANSFORMATIONS
FOR THE TWO-CLASS
PROBLEM

Roger P. Qu, PhD
Department of Biostatistics

MADISON, WISCONSIN
Minimizing Error Probability of Linear Discriminant Analysis Using Box-Cox Transformations for the Two-Class Problem

Roger P. Qu
Department of Biostatistics, University of Wisconsin-Madison,
K6/466, Clinical Sciences Center, 600 Highland Avenue, Madison, WI 53792

ABSTRACT

Searching minimizer of error probability of linear discriminant analysis using Box-Cox transformations to normal approximation is considered for two-class univariate problem. Apparent error rate estimate and bootstrap estimate are used to estimate the true error probability. Numerical comparisons with finite sample simulations are performed. It is found that the approximation to the true minimum value of error probability is satisfactory.

Key words and Phrases: apparent error rate; bootstrap estimate; equal variances; normal approximation.

1 Introduction

Discriminant analysis is a common used statistical method in practice. The most widely used procedures in discriminant analysis in practice are Gaussian linear discriminant rule (LDR) and quadratic discriminant rule (QDR). LDR or QDR are optimal if the underlying distributions are normal. Consequences of violation of normality of the underlying distributions from normality can be severe. One way to overcome this non-normality problem is to transform the underlying distributions to approximate normal before the use of LDR or QDR. A number of people have studied the effect of using transformations. 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. 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.

Both the about work were done under the assumption that the transformed distributions were exact normal. Qu and Loh (1992) considered this problem from the other direction. They 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.

The family of Box-Cox (1964) 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. Qu (1992) considered the performance of LDR and QDR using Box-Cox transformations to approximate normal distributions. His result shows that the use of QDR on the transformed data with unequal variance performs almost uniformly better than LDR with equal variance normal approximation. But as functions of the transformation parameter $\lambda$, the limiting error probabilities using LDR or QDR on the transformed data give rise to two interesting properties:

1. In most cases, there exists a minimizer of the limiting error probabilities either for LDR or QDR;

2. Although the limiting error probabilities using normal approximation with unequal variances are smaller than those using normal approximation with equal variances, the minima of those limiting error rates for LDR and QDR are very close.

If one can estimate the minimizer of the limiting error probabilities, then using Box-Cox transformations with the estimated minimizer may achieve its maximum effect, i.e., reduce the error
rate to its lowest level. From property 2, it may be enough to only consider LDR instead of both LDR and QDR. In this paper, we will consider this problem under the assumption that the two underlying populations are not identical and are positive-valued. We also assume that the second population has the larger population mean. This assumption will reduce our notations significantly without loss of generality.

Two different methods of searching minimizer of error probability will be considered, one with the apparent error rate estimate and the other with bootstrap estimate. Finite sample simulations are performed for the exponential, gamma and lognormal distributions.

2 Apparent error rate estimate of error probabilities

Let \( X_n = (X_1, X_2, \ldots, X_{n_1}) \) and \( Y_n = (Y_1, Y_2, \ldots, Y_{n_2}) \) be two training sets, \( \{X_n\} \) and \( \{Y_n\} \) be two infinite sequences of observations from populations \( F_1 \) and \( F_2 \) respectively. Let \( F_{1,n} \) and \( F_{2,n} \) be the empirical distributions of \( X_n \) and \( Y_n \) respectively. For fixed \( \lambda \), 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 \( t(\lambda) = (EX_\lambda + EY_\lambda)/2 \), and \( t_n(\lambda) = (\bar{X}_\lambda + \bar{Y}_\lambda)/2 \). Denote

\[
p_{1,n}(\lambda, X_n, Y_n) = P[Z_\lambda > (\bar{X}_\lambda + \bar{Y}_\lambda)/2 | Z \sim F_1],
\]

\[
p_{2,n}(\lambda, X_n, Y_n) = P[Z_\lambda < (\bar{X}_\lambda + \bar{Y}_\lambda)/2 | Z \sim F_2],
\]

\[
p_1(\lambda) = P[Z_\lambda > (EX_\lambda + EY_\lambda)/2 | Z \sim F_1],
\]

\[
p_2(\lambda) = P[Z_\lambda < (EX_\lambda + EY_\lambda)/2 | Z \sim F_2].
\]

Let \( p_n(\lambda, X_n, Y_n) = p_{1,n}(\lambda, X_n, Y_n) + p_{2,n}(\lambda, X_n, Y_n) \) and \( p(\lambda) = p_1(\lambda) + p_2(\lambda) \). Obviously, \( p_n(\lambda, X_n, Y_n) \to p(\lambda) \) a.s. for fixed \( \lambda \). In the following, we will show that the convergence is uniform in \( \lambda \).

In the following, we assume that \( F_1 \) and \( F_2 \) have density functions \( f_1 \) and \( f_2 \) with respect to the Lebesgue measure \( m \) on \( R \) respectively and that the expectations \( EX_\lambda \) and \( EY_\lambda \) are finite for \( \lambda \) in an open interval \((a - \delta_0, b + \delta_0)\), where \( a, b \) and \( \delta_0 > 0 \) are constants.
Consider the function

\[ h(\lambda, t) = \begin{cases} 
(\lambda t + 1)^{1/\lambda}, & \text{if } \lambda \neq 0 \text{ and } \lambda t + 1 > 0, \\
\exp(t), & \lambda = 0.
\end{cases} \]

For \( \lambda \neq 0 \), \((\lambda t + 1)^{1/\lambda}\) is continuous everywhere on its domain \( D \). As \( \lambda \to 0 \),

\[ \lim_{\lambda \to 0} h(\lambda, t) = \lim_{\lambda \to 0} \exp\{\log(\lambda t + 1)/\lambda\} = \exp(t) = h(0, t). \]

Thus \( h(\lambda, t) \) is continuous in \((\lambda, t)\) on its domain.

**Lemma 1** Let \( g(\lambda) \) be a function and \( g_n(\lambda) \) be a sequence of functions of \( \lambda \in [a, b] \). Suppose that \( g(\lambda) \) is continuous in \( \lambda \). Let \( g(\lambda) \) and \( g_n(\lambda) \), \( n = 1, 2, \ldots \), be such that \((\lambda, g(\lambda))\) and \((\lambda, g_n(\lambda))\) are in the domain \( D \) of the function \( h(\lambda, t) \), and that \( g_n(\lambda) \to g(\lambda) \) uniformly in \( \lambda \in [a, b] \). Then \( h(\lambda, g_n(\lambda)) \to h(\lambda, g(\lambda)) \) uniformly in \( \lambda \in [a, b] \).

**Proof.** Since \( g(\lambda) \) is continuous in \( \lambda \), the set \( A = \{(\lambda, g(\lambda)); \lambda \in [a, b]\} \) is compact in \( R^2 \). Note that the domain \( D = \{(\lambda, t); \lambda t + 1 > 0\} \) of \( h(\lambda, t) \) is open in \( R^2 \) and \( A \subset D \). Let \( ||(\lambda, t)|| = (\lambda^2 + t^2)^{1/2} \) be the norm of \((\lambda, t)\) in \( R^2 \). Let \( B' \) denote the boundary of a set \( B \) in \( R^2 \). Since \( A \) is compact, there is an open set \( A_1 \) such that \( A \subset A_1 \subset D \), and the distance \( d(A_1', A) = \inf\{||(\lambda - \lambda_1, t - t_1)||; (\lambda, t) \in A, (\lambda_1, t_1) \in A_1'\} \) is positive. Let \( \bar{A}_1 \) be the closure of \( A_1 \). Then \( \bar{A}_1 \subset D \) and \( \bar{A}_1 \) is compact. Thus \( h(\lambda, t) \) is uniformly continuous in \((\lambda, t)\) on \( \bar{A}_1 \). So for any \( \varepsilon > 0 \), there is a \( \delta < d(A_1', A) \) such that \( |h(\lambda, t) - h(\lambda_1, t_1)| \leq \varepsilon \) if \((\lambda, t)\) and \((\lambda_1, t_1)\) are in \( \bar{A}_1 \) and \( ||(\lambda - \lambda_1, t - t_1)|| \leq \delta \).

Since \( g_n(\lambda) \to g(\lambda) \) uniformly on \([a, b]\), there is \( N \) such that \( n \geq N \) implies that \( |g_n(\lambda) - g(\lambda)| = ||(\lambda - \lambda, g_n(\lambda) - g(\lambda)|| \leq \delta \) for all \( \lambda \in [a, b] \). Thus \((\lambda, g_n(\lambda)) \in \bar{A}_1 \) for all \( \lambda \in [a, b] \) and \( n \geq N \) since \((\lambda, g(\lambda)) \in A \) for \( \lambda \in [a, b] \). So \( |h(\lambda, g_n(\lambda)) - h(\lambda, g(\lambda))| \leq \varepsilon \) for \( n \geq N \) and \( \lambda \in [a, b] \). This finishes the proof. \( \square \)

**Corollary 1** \( h(\lambda, t_n(\lambda)) \to h(\lambda, t(\lambda)) \) uniformly in \( \lambda \in [a, b] \) a.s.

**Proof.** By Rubin’s (1956) Theorem, \( t_n(\lambda) \to t(\lambda) \) uniformly in \( \lambda \in [a, b] \) a.s. So the result is an immediate result of Lemma 1. \( \square \)

**Theorem 1** \( p_n(\lambda, X_n, Y_n) \to p(\lambda) \) uniformly in \( \lambda \in [a, b] \) a.s.
Proof. For any $\varepsilon > 0$, let $\delta$ be such that for measurable set $A$, $m(A) \leq \delta$ implies that $\int_A |f_1| dm \leq \varepsilon$, see Royden (1968) Proposition 3.13. By Lemma 1, there is a $N$ be such that $n \geq N$ implies $|h(\lambda, t_n(\lambda)) - h(\lambda, t(\lambda))| \leq \delta/2$ for all $\lambda \in [a, b]$. Then when $n \geq N$,

$$\left| p_{1,n}(\lambda, X_n, Y_n) - p_1(\lambda) \right| = \left| P[Z > h(\lambda, t_n(\lambda))|Z \sim F_1] - P[Z > h(\lambda, t(\lambda))|Z \sim F_1] \right| \leq P[h(\lambda, t(\lambda)) - \delta/2 < Z < h(\lambda, t(\lambda)) + \delta/2|Z \sim F_1] \leq \varepsilon.$$ 

A similar argument holds for $p_{2,n}(\lambda, X_n, Y_n)$. Hence the claim is true. $\Box$

Now we consider the estimation of $p_n(\lambda, X_n, Y_n)$. The reason to consider this quantity instead of $p(\lambda)$ is that the actual error rate given the observations $X_n$ and $Y_n$ is $p_n(\lambda, X_n, Y_n)$. Thus it is natural to consider $p_n(\lambda, X_n, Y_n)$. An obvious estimate of $p_{1,n}(\lambda, X_n, Y_n)$ is the so called apparent error rate

$$p_{1,n}^a(\lambda, X_n, Y_n) = n_1^{-1} \sum_{i=1}^{n_1} I[X_{\lambda,i} > (\bar{X}_\lambda + \bar{Y}_\lambda)/2] = P[Z > h(\lambda, t_n(\lambda))|Z \sim F_{1,n}].$$

Similarly, the apparent error rate estimate of $p_{2,n}(\lambda, X_n, Y_n)$ is

$$p_{2,n}^a(\lambda, X_n, Y_n) = n_2^{-1} \sum_{i=1}^{n_2} I[Y_{\lambda,i} < (\bar{X}_\lambda + \bar{Y}_\lambda)/2] = P[Z < h(\lambda, t_n(\lambda))|Z \sim F_{2,n}].$$

There is another interpretation of $p_{1,n}^a(\lambda, X_n, Y_n)$. Actually, it is the estimate of $p_1(\lambda)$ using $F_{1,n}$ and $F_{2,n}$ as estimates of $F_1$ and $F_2$. Denote $p_n^a(\lambda, X_n, Y_n) = p_{1,n}^a(\lambda, X_n, Y_n) + p_{2,n}^a(\lambda, X_n, Y_n)$. The relationships among $p(\lambda)$, $p_n(\lambda, X_n, Y_n)$ and $p_n^a(\lambda, X_n, Y_n)$ are reflected in the following theorem.

**Theorem 2** (a). $p_n(\lambda, X_n, Y_n) - p_n^a(\lambda, X_n, Y_n) \rightarrow 0$ uniformly in $\lambda \in [a, b]$ a.s. (b). $p_n^a(\lambda, X_n, Y_n) \rightarrow p(\lambda)$ uniformly in $\lambda \in [a, b]$ a.s.

Proof. Part (b) is an immediate result of part (a) because of Theorem 1. Thus it is enough to prove part (a). In fact, by Dvoretzky-Kiefer-Wolfowitz's Theorem, see Serfling (1980),

$$|p_{1,n}(\lambda, X_n, Y_n) - p_{1,n}^a(\lambda, X_n, Y_n)| = |F_{1,n}(h(\lambda, t_n(\lambda))) - F_1(h(\lambda, t_n(\lambda)))| \leq \sup_{x \in \mathbb{R}} |F_{1,n}(x) - F_1(x)| \rightarrow 0$$

a.s. $\Box$
Suppose that \( p(\lambda) \) has unique minimizer \( \lambda_0 \) over \([a, b]\). Let \( \hat{\lambda} \) and \( \hat{\lambda}^a \) be the minimizers of \( p_n(\lambda, X_n, Y_n) \) and \( p_n^a(\lambda, X_n, Y_n) \) respectively. Then by Lemma 2 of Chen 1990, we have

**Corollary 2** (a) \( \hat{\lambda} - \hat{\lambda}^a \to 0 \ a.s. \) (b) \( \hat{\lambda} \to \lambda_0 \ a.s. \) and \( \hat{\lambda}^a \to \lambda_0 \ a.s. \)

**Remark.** The unknown quantity \( \hat{\lambda} \) is the one we are interested in. Its estimate \( \hat{\lambda}^a \) is obtained from \( p_n^a(\lambda, X_n, Y_n) \). Corollary 2 tells us that if \( \lambda_0 \) exists, then the difference between \( \hat{\lambda} \) and \( \hat{\lambda}^a \), \( \hat{\lambda} - \hat{\lambda}^a \to 0 \ a.s. \) and both of them converge to \( \lambda_0 \).

## 3 Bootstrap estimate of error probabilities

The apparent error rate estimate in the previous section has the merit of being simple. But as an estimate of the true error rate \( p_n(\lambda, X_n, Y_n) \), it has been often criticized as too "optimistic" since it uses the same data in the classifier to construct the estimate of the true error rate. Various methods have been proposed to correct the bias using \( p_n^a(\lambda, X_n, Y_n) \). One of them is the so called "bootstrap" estimate. Rewrite

\[
p_{1,n}(\lambda, X_n, Y_n) = p_{1,n}(\lambda, X_n, Y_n) - p_{1,n}^a(\lambda, X_n, Y_n) + p_{1,n}^a(\lambda, X_n, Y_n) \]

\[
= \gamma_{1,n}(\lambda, X_n, Y_n) + p_{1,n}^a(\lambda, X_n, Y_n),
\]

where \( \gamma_{1,n}(\lambda, X_n, Y_n) = p_{1,n}(\lambda, X_n, Y_n) - p_{1,n}^a(\lambda, X_n, Y_n) \) is the bias using the apparent error rate as the estimate of true error rate. Since the true error rate \( p_{1,n}(\lambda, X_n, Y_n) \) is unknown, so is \( \gamma_{1,n}(\lambda, X_n, Y_n) \). But if we can somehow estimate this bias, then we can improve our estimation. Let \( \gamma_{1,n,e}(\lambda) = E[\gamma_{1,n}(\lambda, X_n, Y_n)] \). If \( \gamma_{1,n,e}(\lambda) \) were known, we could use it as the estimate of \( \gamma_{1,n}(\lambda, X_n, Y_n) \). Of course, \( \gamma_{1,n,e}(\lambda) \) is unknown since the distributions of \( X_n \) and \( Y_n \) are unknown. Thus we need to estimate the quantity \( \gamma_{1,n,e}(\lambda) \). The bootstrap method can be used to achieve this. References about the bootstrap method can be found in Efron (1982).

The bootstrap is essentially a resampling method. Let \( X_n^* = (X_1^*, X_2^*, \ldots, X_{n_1}^*) \) and \( Y_n^* = (Y_1^*, Y_2^*, \ldots, Y_{n_2}^*) \) be two samples from the empirical distributions \( F_{1,n} \) and \( F_{2,n} \) respectively. Let \( F_{1,n}^a \) and \( F_{2,n}^a \) be the empirical distributions of \( X_n^* \) and \( Y_n^* \) respectively. Let \( X_\lambda^* \) and \( \bar{Y}_\lambda^* \) be the counterparts of \( X_\lambda \) and \( \bar{Y}_\lambda \) for the bootstrap samples \( X_n^* \) and \( Y_n^* \). Denote \( t_n^*(\lambda) = (X_\lambda^* + \bar{Y}_\lambda^*)/2 \). The counterparts of \( p_{1,n}(\lambda, X_n, Y_n) \), \( p_{1,n}^a(\lambda, X_n, Y_n) \) and \( \gamma_{1,n}(\lambda, X_n, Y_n) \) are, respectively

\[
p_{1,n}^*(\lambda, X_n^*, Y_n^*) = P[Z > h(\lambda, t_n^*(\lambda))|Z ~ F_{1,n}],
\]
\begin{align*}
p^{*}_{1, n}(\lambda, X^*_n, Y^*_n) &= n_1^{-1} \sum_{i=1}^{n_1} I[ X^*_i > h(\lambda, t^*_n(\lambda))], \\
\gamma^*_1(\lambda, X^*_n, Y^*_n) &= p^{*}_{1, n}(\lambda, X^*_n, Y^*_n) - p^{*}_{0, n}(\lambda, X^*_n, Y^*_n).
\end{align*}

Let \( P^* \) denote probability under bootstrap sampling and \( E^* \) denote expectation calculated under \( P^* \). Then the bootstrap estimate of \( \gamma_{1, n, e}(\lambda) \) is the expectation of \( \gamma^*_1(\lambda, X^*_n, Y^*_n) \) calculated under the probability \( P^* \)

\[ \gamma^*_1(\lambda, X^*_n, Y^*_n) = E^* \gamma^*_1(\lambda, X^*_n, Y^*_n). \]

In practice, \( \gamma^*_1(\lambda, X^*_n, Y^*_n) \) is calculated by simulation. Independent bootstrap samples \( (X^*_n)^{(1)}(Y^*_n)^{(1)}, (X^*_n)^{(2)}(Y^*_n)^{(2)}, \ldots, (X^*_n)^{(B)}(Y^*_n)^{(B)} \) are generated, and for each \( b = 1, 2, \ldots, B \), \( \gamma^*_1(\lambda, X^*_n, Y^*_n) \) is calculated. The average of the \( B \) realizations of \( \gamma^*_1(\lambda, X^*_n, Y^*_n) \) serves as the expectation of \( \gamma^*_1(\lambda, X^*_n, Y^*_n) \). With today's highly developed computers, the calculation of \( \gamma^*_1(\lambda, X^*_n, Y^*_n) \) is not hard to be accomplished.

Now we investigate the large sample properties of the bootstrap estimate

\[ p^*_1(\lambda, X_n, Y_n) = \gamma^*_1(\lambda, X_n, Y_n) + p^{*}_{0, n}(\lambda, X_n, Y_n). \]

Our goal is to establish the uniform convergence of \( p^*_1(\lambda, X_n, Y_n) \) to its limit \( p_1(\lambda) \). We will use "a.s. \( P^* \)" to denote the almost sure property under the bootstrap probability \( P^* \).

**Lemma 2** For almost all sequences \( \{X_n\} \), \( F^*_1 \rightarrow F_1 \) weakly a.s. \( P^* \).

**Proof.** Denote \( D^*_n = \sup_{x \in R} |F^*_1(x) - F_{1, n}(x)| \). Then by Dvoretzky-Kiefer-Wolfowitz's Theorem, for \( d > 0 \),

\[ P^*(D^*_n \geq d) \leq C \exp\{-2nd^2\}. \]

Thus for any \( \varepsilon > 0 \),

\[ \sum_{i=1}^{\infty} P^*(D^*_n > \varepsilon) \leq C \sum_{i=1}^{\infty} \exp\{-2n\varepsilon^2\} < \infty. \]

By the first Borel-Cantelli Lemma, see Billingsley 1986, \( P^*(D^*_n > \varepsilon \ i.o.) = 0 \), where i.o. represents "infinitely often". Therefore \( D^*_n \rightarrow 0 \) a.s. \( P^* \). Note that \( F_{1, n} \rightarrow F_1 \) weakly for almost all sequences \( \{X_n\} \). This finishes the proof. \( \square \)

**Lemma 3** For almost all sequences \( \{X_n\} \) and \( \{Y_n\} \), \( \bar{X}^*_\lambda = n_1^{-1} \sum_{i=1}^{n_1} X^*_{\lambda, i} \rightarrow EX_\lambda \) and \( \bar{Y}^*_\lambda = n_2^{-1} \sum_{i=1}^{n_2} Y^*_{\lambda, i} \rightarrow EY_\lambda \) uniformly in \( \lambda \in [a, b] \) a.s. \( P^* \).
Proof. For \( d_0 > 1 \), denote \( T_N^0 = N^{-1} \sum_{i=1}^{N} X_i^* I(X_i^* \geq d_0) \). Then by Csörgö (1992), \( T_N^0 \to E X_\lambda I(X \geq d_0) \) a.s. Similarly to the proof of Lemma 7 of Qu 1992, we have

\[
\sup_{N \geq 1} N^{-1} \sum_{i=1}^{N} X_i^* I(X_i^* \geq d) \to 0 \text{ as } d \to 0 \text{ a.s.} P^*
\]

for almost all sequences \( \{X_n\} \). Since \( F_{1,n}^* \to F_1 \text{ a.s.} P^* \) for almost all sequences \( \{X_n\} \), the rest of the proof is similar to that of Theorem 3 of Qu 1992. \( \square \)

**Corollary 3** For almost all sequences \( \{X_n\} \) and \( \{Y_n\} \), \( h(\lambda, t_n^*(\lambda)) \to h(\lambda, t(\lambda)) \) uniformly in \( \lambda \in [a, b] \) a.s.\( P^* \).

**Corollary 4** For almost all sequences \( \{X_n\} \) and \( \{Y_n\} \), \( p_{1,n}^*(\lambda, X_n^*, Y_n^*) \to p_1(\lambda) \) and \( p_{1,n}^{a,*}(\lambda, X_n^*, Y_n^*) \to p_1(\lambda) \) uniformly in \( \lambda \in [a, b] \) a.s.\( P^* \).

Proof. For the first part,

\[
|p_{1,n}^*(\lambda, X_n^*, Y_n^*) - p_1(\lambda)|
\]

\[
\leq |p_{1,n}^*(\lambda, X_n^*, Y_n^*) - P[Z > h(\lambda, t_n^*(\lambda))]| Z \sim F_1|
\]

\[
+ |P[Z > h(\lambda, t_n^*(\lambda))]| Z \sim F_1| - p_1(\lambda)|.
\]

The first term on the right side is \( \leq D_{1,n} = \sup_{x \in R} |F_{1,n}(x) - F_1(x)| \). So it converges to 0 by Dvoretzky-Kiefer-Wolfowitz's Theorem. For any \( \varepsilon > 0 \), there is \( N \) such that \( n \geq N \) implies

\[
|h(\lambda, t_n^*(\lambda)) - h(\lambda, t(\lambda))| \leq \varepsilon \text{ for all } \lambda \in [a, b].
\]

Thus for \( n \geq N \), the second term on the right side is

\[
\leq P[h(\lambda, t(\lambda)) - \varepsilon < Z < h(\lambda, t(\lambda)) + \varepsilon] Z \sim F_1].
\]

Since \( \varepsilon \) is arbitrary, the second term on the right side converges to 0 uniformly in \( \lambda \in [a, b] \) a.s.\( P^* \) by Royden (1968) Proposition 3.13.

Consider the second part of the Corollary. By Lemma 2 and Corollary 3, for any \( \varepsilon > 0 \), there exists a \( N \), such that \( n > N \) implies that

\[
\sup_{\lambda \in [a, b]} |h(\lambda, t_n^*(\lambda)) - h(\lambda, t(\lambda))| < \varepsilon \text{ and } D_{1,n}^{a,*} = \sup_{x \in R} |F_{1,n}^*(x) - F_1(x)| < \varepsilon.
\]

So for \( n \geq N \),

\[
|p_{1,n}^{a,*}(\lambda, X_n^*, Y_n^*) - p_1(\lambda)|
\]

\[
\leq |p_{1,n}^{a,*}(\lambda, X_n^*, Y_n^*) - F_{1,n}(h(\lambda, t_n^*(\lambda)))|.
\]
\[ + |F_{1,n}(h(\lambda, t^*_n(\lambda))) - F_1(h(\lambda, t^*_n(\lambda)))| + |F_1(h(\lambda, t^*_n(\lambda))) - p_1(\lambda)| \]
\[ \leq D_{1,n}^* + D_{1,n} + [F_1(h(\lambda, t(\lambda)) + \varepsilon) - F_1(h(\lambda, t(\lambda)) - \varepsilon)]. \]

Using Dvoretzky-Kiefer-Wolfowitz's Theorem and Royden (1968) Proposition 3.13 finishes the proof. \qed

**Corollary 5** For almost all sequences \( \{X_n\} \) and \( \{Y_n\} \), \( E^* p_{1,n}^*(\lambda, X_n^*, Y_n^*) \to p_1(\lambda) \), and \( E^* p_{1,n}^{a,*}(\lambda, X_n^*, Y_n^*) \to p_1(\lambda) \) uniformly in \( \lambda \in [a, b] \).

**Proof.** By Corollary 4,
\[
\sup_{\lambda \in [a, b]} |p_{1,n}^*(\lambda, X_n^*, Y_n^*) - p_1(\lambda)| \to 0 \text{ a.s. } P^*.
\]
By Lebesgue dominated convergence theorem,
\[
E^* \sup_{\lambda \in [a, b]} |p_{1,n}^*(\lambda, X_n^*, Y_n^*) - p_1(\lambda)| \to 0.
\]
So
\[
\sup_{\lambda \in [a, b]} |E^* p_{1,n}^*(\lambda, X_n^*, Y_n^*) - p_1(\lambda)| \leq E^* \sup_{\lambda \in [a, b]} |p_{1,n}^*(\lambda, X_n^*, Y_n^*) - p_1(\lambda)| \to 0.
\]
A similar argument holds for the second part. \qed

Let \( p_{2,n}^a(\lambda, X_n^*, Y_n^*) \), \( p_{2,n}^{a,*}(\lambda, X_n^*, Y_n^*) \), \( \gamma_{2,n,e}^a(\lambda, X_n, Y_n) \) and \( p_{2,n}^b(\lambda, X_n, Y_n) \) be the counterparts of \( p_{1,n}^*(\lambda, X_n^*, Y_n^*) \), \( p_{1,n}^{a,*}(\lambda, X_n^*, Y_n^*) \), \( \gamma_{1,n,e}^a(\lambda, X_n, Y_n) \) and \( p_{1,n}^a(\lambda, X_n, Y_n) \). All the above results related to the first error rate \( p_{1,n}(\lambda, X_n, Y_n) \) are true for their counterparts for the second error rate \( p_{2,n}(\lambda, X_n, Y_n) \). Let \( p_n^a(\lambda, X_n, Y_n) = p_{1,n}^b(\lambda, X_n, Y_n) + p_{2,n}^b(\lambda, X_n, Y_n) \) be the bootstrap estimate of the true error rate \( p_n(\lambda, X_n, Y_n) \).

**Theorem 3** \( p_n^b(\lambda, X_n, Y_n) \to p(\lambda) \) uniformly in \( \lambda \in [a, b] \) a.s.

Let \( \hat{\lambda}^b \) be the minimizer of \( p_n^b(\lambda, X_n, Y_n) \).

**Corollary 6** (a). \( \hat{\lambda}^b - \lambda \to 0 \) a.s., (b). \( \hat{\lambda}^b \to \hat{\lambda}_0 \) a.s.
4 Finite Sample Simulation

To see the performance of the approximations, finite sample simulations were conducted for the gamma, exponential and lognormal distributions. The two distributions $F_1$ and $F_2$ in the experiment were chosen different by a location shift $c$. We use $p_r$ to denote the error probability on the raw data; $p_t$ the true minimum value of error probability as a function of $\lambda$ for the transformed data; $p_a$ the minimum error probability using apparent error rate estimate and $p_b$ the minimum error probability using bootstrap estimate; $d_a = |p_t - p_a|$ and $d_b = |p_t - p_b|$ are the differences between the true and estimated minimum error probabilities; $s_a$ and $s_b$ are the standard errors of $p_a$ and $p_b$ respectively. The experiment was done for sample sizes $n_1 = n_2 = 20, 30, 50$. The number of trails is 400 in each case. The number of iterations for bootstrapping is 100. For each trail, samples $X_1, X_2, \ldots, X_{n_1}$ and $Y_1, Y_2, \ldots, Y_{n_1}$ were obtained, then $p_r, p_a(\lambda, X_n, Y_n), p_a^b(\lambda, X_n, Y_n)$ and $p_b(\lambda, X_n, Y_n)$ were calculated, as functions of $\lambda$, and the minimum error probability of each method was found. The result was the average of the 400 trails for each method, e.g., $p_r = 400^{-1} \sum_{i=1}^{400} p_{r,i}$, where $p_{r,i}$ was the true error probability without transformation for the $i$-th trail.

The simulation results are consistent with what we expected. As we can see, bootstrap estimate does improve the estimation by lifting the apparent error rate estimate which usually underestimates the true error probability. Also, in most cases, the absolute difference $d_b$ between $p_b$ and $p_t$ is smaller than the absolute difference $d_a$ between $p_a$ and $p_t$. This indicates that not only bootstrap estimate is better on average, but also on the individual based comparison.

Table 1 Finite sample simulations for the exponential distribution with location shift $c$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$p_r$</th>
<th>$p_t$</th>
<th>$p_a$</th>
<th>$s_a$</th>
<th>$d_a$</th>
<th>$p_b$</th>
<th>$s_b$</th>
<th>$d_b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>exp(1), $c = .3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>.888</td>
<td>.766</td>
<td>.686</td>
<td>.126</td>
<td>.114</td>
<td>.738</td>
<td>.141</td>
<td>.109</td>
</tr>
<tr>
<td>30</td>
<td>.888</td>
<td>.761</td>
<td>.698</td>
<td>.105</td>
<td>.094</td>
<td>.740</td>
<td>.114</td>
<td>.088</td>
</tr>
<tr>
<td>50</td>
<td>.889</td>
<td>.758</td>
<td>.717</td>
<td>.084</td>
<td>.070</td>
<td>.746</td>
<td>.088</td>
<td>.067</td>
</tr>
<tr>
<td>exp(1), $c = 1.2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>.527</td>
<td>.304</td>
<td>.288</td>
<td>.097</td>
<td>.077</td>
<td>.312</td>
<td>.108</td>
<td>.086</td>
</tr>
<tr>
<td>30</td>
<td>.527</td>
<td>.304</td>
<td>.295</td>
<td>.085</td>
<td>.070</td>
<td>.311</td>
<td>.091</td>
<td>.073</td>
</tr>
<tr>
<td>50</td>
<td>.532</td>
<td>.304</td>
<td>.299</td>
<td>.065</td>
<td>.052</td>
<td>.309</td>
<td>.068</td>
<td>.054</td>
</tr>
</tbody>
</table>
Table 2  Finite sample simulations for the gamma distribution with location shift $c$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$p_r$</th>
<th>$p_l$</th>
<th>$p_a$</th>
<th>$s_a$</th>
<th>$d_a$</th>
<th>$p_b$</th>
<th>$s_b$</th>
<th>$d_b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>gamma(3,1), $c = 1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>.779</td>
<td>.738</td>
<td>.642</td>
<td>.139</td>
<td>.138</td>
<td>.672</td>
<td>.150</td>
<td>.133</td>
</tr>
<tr>
<td>30</td>
<td>.778</td>
<td>.737</td>
<td>.676</td>
<td>.119</td>
<td>.107</td>
<td>.696</td>
<td>.124</td>
<td>.105</td>
</tr>
<tr>
<td>50</td>
<td>.777</td>
<td>.736</td>
<td>.693</td>
<td>.092</td>
<td>.080</td>
<td>.705</td>
<td>.094</td>
<td>.079</td>
</tr>
<tr>
<td>gamma(3,1), $c = 3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>.371</td>
<td>.316</td>
<td>.253</td>
<td>.098</td>
<td>.096</td>
<td>.267</td>
<td>.106</td>
<td>.096</td>
</tr>
<tr>
<td>30</td>
<td>.369</td>
<td>.316</td>
<td>.266</td>
<td>.084</td>
<td>.080</td>
<td>.277</td>
<td>.089</td>
<td>.079</td>
</tr>
<tr>
<td>50</td>
<td>.367</td>
<td>.316</td>
<td>.284</td>
<td>.063</td>
<td>.057</td>
<td>.292</td>
<td>.066</td>
<td>.057</td>
</tr>
</tbody>
</table>

Table 3  Finite sample simulations for the lognormal distribution with location shift $c$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$p_r$</th>
<th>$p_l$</th>
<th>$p_a$</th>
<th>$s_a$</th>
<th>$d_a$</th>
<th>$p_b$</th>
<th>$s_b$</th>
<th>$d_b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>lognorm(0,1), $c = .3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>.933</td>
<td>.837</td>
<td>.748</td>
<td>.146</td>
<td>.137</td>
<td>.782</td>
<td>.152</td>
<td>.128</td>
</tr>
<tr>
<td>30</td>
<td>.934</td>
<td>.834</td>
<td>.779</td>
<td>.115</td>
<td>.105</td>
<td>.806</td>
<td>.121</td>
<td>.103</td>
</tr>
<tr>
<td>50</td>
<td>.935</td>
<td>.833</td>
<td>.795</td>
<td>.096</td>
<td>.080</td>
<td>.810</td>
<td>.100</td>
<td>.076</td>
</tr>
<tr>
<td>lognorm(3,2), $c = 10$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>.953</td>
<td>.726</td>
<td>.666</td>
<td>.149</td>
<td>.110</td>
<td>.715</td>
<td>.156</td>
<td>.105</td>
</tr>
<tr>
<td>30</td>
<td>.955</td>
<td>.716</td>
<td>.680</td>
<td>.121</td>
<td>.084</td>
<td>.721</td>
<td>.127</td>
<td>.085</td>
</tr>
<tr>
<td>50</td>
<td>.957</td>
<td>.710</td>
<td>.690</td>
<td>.094</td>
<td>.064</td>
<td>.719</td>
<td>.095</td>
<td>.068</td>
</tr>
</tbody>
</table>

References


14, 1-7.


quadratic discriminant function to certain types of non-normality, *Commun. Statist.,* 1, 39-56.


for the two-class problem.

the two-class problem, *Commun. Statist. – Th. and Meth.,* A21, 2757–2774.

