Nonparametric Estimation of the Threshold at an Operating Point on the ROC Curve

By

Waleed Ahmed Yousef

D.Sc. in Electrical and Computer Engineering, January 2006, The George Washington University

M.Sc. in Computer Science, November 1999, Helwan University

B.Sc. in Electrical Engineering, June 1995, Ain Shams University

A Thesis submitted to

The Faculty of

The Columbian College of Arts and Sciences

of The George Washington University

in partial satisfaction of the requirements for the degree of

Master of Science

January, 30, 2007

Thesis directed by

Subrata Kundu,

Professor of Statistics

Robert F. Wagner,

Senior Biomedical Research Scientist at CDRH, FDA.

Abstract

In the problem of binary classification or diagnostic testing, the classification algorithm or diagnostic test produces a continuous decision variable which is compared to a critical value (or threshold). Test values above (or below) that threshold are called positive (or negative) for disease. There are two types of errors at every threshold value. The relationship between these two types is the receiver operating characteristic curve (ROC). The present work is concerned with the inverse problem; i.e., given the ROC curve (or its estimate) of a particular classification rule, what is the value of the threshold ξ that leads to a specific operating point on that curve, i.e., a specific pair of the two types of error. In the present work we assume the availability of a finite sample from each distribution (class), and no knowledge about the two distributions; i.e., the threshold is estimated nonparametrically. Asymptotic distribution is derived for the proposed estimator. Simulation results are presented for finite sample size. Finding a particular threshold value is crucial in medical diagnoses, among other fields, where a medical test is used to classify a patient as "diseased" or "nondiseased" based on comparing the test result to a particular threshold value. When the ROC is estimated, an operating point is obtained by selecting one type of errors, and get the other one from the estimated curve. Threshold estimation can then be viewed as a quantile estimation for one distribution but with the utilization of the second one.

Acknowledgements

All praises are due to Allah, Most Gracious, Most Merciful, Who said: *"When the Earth is shaken to its (utmost) convulsion. And the Earth throws up its burdens (from within). And man cries (distressed); 'What is the matter with it?'. On that Day will it declare its tidings. For that thy Lord will have given it inspiration. On that Day will people proceed in companies sorted out, to be shown the Deeds that they (had done). Then shall anyone who has done an atom's weight of good, see it! And anyone who has done an atom's weight of evil, shall see it."*[1]

I want to express my thanks to my supervisors, Doctors Subrata Kundu and Robert Wagner for their time, assistance, and support.

I would like to express my special thanks and acknowledgement to many people at the Columbian School. Professor Michael Moses, the Associate Dean for Graduate Studies, showed his great cooperation to allow me to join the program. Professor Tapan Nayak, the head of Statistics Department, and Professor Efstathia Bura were very supportive. Professor Hosam Mahmoud was very generous in affording voluntary many hours of discussion.

This project was supported in part by a CDRH Medical Device Fellowship, granted to Waleed Yousef, and administered by the Oak Ridge Institute for Science and Education (ORISE).

[1] The Holy Qur'ān, 99

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Consider the binary classification problem, in which a multidimensional random vector X belongs to one of two classes, ω_1 or ω_0; we then have two distributions P_1^X and P_0^X, with densities p_1 and p_0 respectively. Consider a decision function $\phi(X)$ that decides whether an observation x belongs to ω_1 or ω_0 based on a decision value—e.g., a medical test—$\eta(X)$ whether it is larger or smaller than a particular threshold value ξ. This nonrandomized rule is given by

$$\phi(X) = \begin{cases} 1, & \text{when} \quad \eta(X) < \xi; \\ \gamma, & \text{when} \quad \eta(X) = \xi; \\ 0, & \text{when} \quad \eta(X) > \xi. \end{cases}$$

Since X is distributed either as P_1 or P_0, according to whether the true class is ω_1 or ω_0, then $\eta(X)$ will have corresponding distributions denoted as F or G respectively. Without loss of generality we can assume that—as it is customary in many applications, e.g., medical diagnoses—the decision values under the null hypothesis G are "in general" larger than those under the alternative F; a special case, which is not necessary for our analysis, would be if G is stochastically larger than F. Therefore, rejecting the null hypothesis, i.e., classifying the observation as belonging to ω_1, will be made for small values of the decision values $\eta(X)$.

At every chosen threshold ξ there will be two different types of errors. Type I error, $e_1 = \mathrm{E}_0\,\phi$ is the probability of rejecting the null while it is true; also called level-α test, where $\alpha = e_1$. Type II error, $e_2 = 1 - \mathrm{E}_1\,\phi$, is the probability of rejecting the alternative while it is true. The power of this test, $\beta = \mathrm{E}_1\,\phi$, is desired to be maximized for a particular type I error. If we choose $\eta(X) = p_0(X)/p_1(X)$ this will be the most powerful level-α test, which is a manifestation of the

famous Neyman-Pearson lemma (see Lehmann and Romano, 2005, Ch. 3). Figure 1.1 illustrates the two densities under the two hypotheses along with the two types of errors

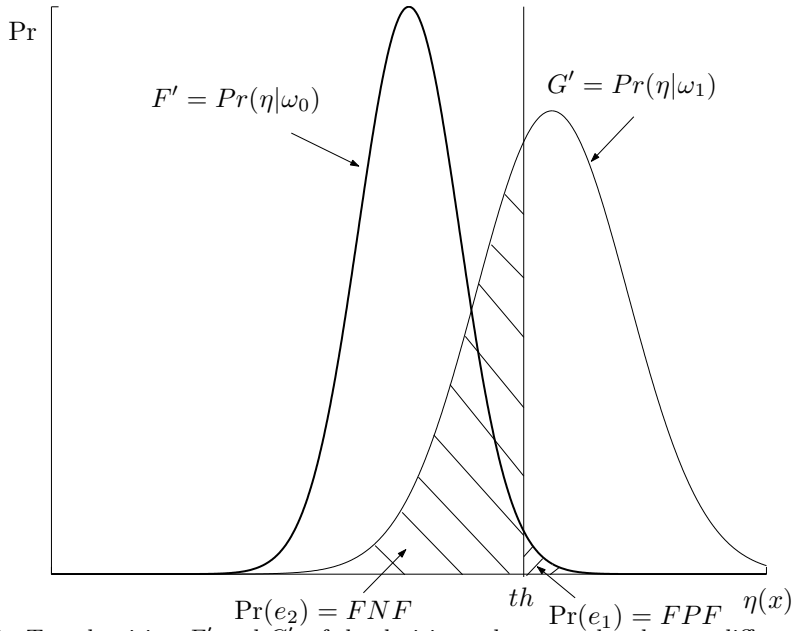


Figure 1.1. Two densities, F' and G', of the decision values η under the two different hypothesis. The two types of error are indicated.

The Ordinal Dominance Curve (ODC) is a plot of β vs. α at every possible threshold value ξ. In terms of G and F, this is a plot of $F(\xi)$ vs. $G(\xi)$, $\forall\ \xi$; equivalently it is a plot of $FG^{-1}(p)$ vs. p, $0 \leq p \leq 1$. In some applications, e.g., Automatic Target Recognition (ATR), and imaging and medical diagnoses, it is customary to plot $1-\alpha$ vs. $1-\beta$. The resulting curve is called the Receiver Operating Characteristic (ROC) curve. In that context, $1-\alpha$ is called the True Positive Fraction (TPF) and $1-\beta$ is called the False Positive Fraction (FPF). The solid curve in Figure 2.1 is an illustration for an ODC.

The costs of the two types of error are typically not equal. For example, missing a cancer (a false negative) greatly outweighs sending a nondiseased patient to biopsy (a false positive). The prior probabilities π_1 and π_0 of the two classes ω_1 and ω_0 also enter into the choice of the optimum threshold setting. If the decision function η is the likelihood ratio p_0/p_1 and c_{ij} is the cost of deciding i while the truth is j, then selecting ξ to be equal to $c_{01}\pi_1/c_{10}\pi_0$ gives rise to the Bayes decision

rule which minimizes the risk (see Anderson, 2003, Ch. 6). In general, η is not the likelihood ratio and we need to know which threshold value ξ would give a particular operating point on the ODC, i.e., α and β, such that the risk will be minimized. If we do not know the costs and priors, the designers of the classification rule would still like to operate at a particular operating point that they determine to satisfy some subjective criteria. In either case, the decision maker must address the same question, i.e., what is the value of ξ that achieves the operating point $(p, s) = (G(\xi), F(\xi))$ on the ODC?

The answer is straightforward if we know the distributions G and F. However, when we do not know these distributions we construct the classification rule with the aid of a "training set" $\mathbf{tr} = \{x_i \in \omega_0,\ i = 1, \cdots, n\} \cup \{x_j \in \omega_1,\ j = 1, \cdots, m\}$. The resulting decision function is $\eta_{\mathbf{tr}}$. Several parametric and nonparametric techniques are available in the literature to construct a classification rule, which is not the target of the present work. Estimating the ODC (or ROC) of $\eta_{\mathbf{tr}}$ is of great interest for ROC analysis. Several approaches are available in the literature for that task. It is not our intention here to give a full account of the history of ROC estimation/analysis; however, some key papers can be cited (Dorfman and Alf, —1969—; Hsieh and Turnbull, —1996—; Metz and Pan, —1999—; Pepe, —2000—; Qin and Zhang, —2003—).

A naive estimator of the ODC is the unsmooth empirical ODC, which we denote by $\widehat{ODC}$; see Figure 2.1. It is a plot of the empirical distribution function $F_m(\xi)$ vs. $G_n(\xi)$ at every threshold value ξ (see Section 2). The empirical ODC has several attractive distribution-free features. It converges to the true ODC, i.e., $FG^{-1}(p)$, $\forall\ 0 \leq p \leq 1$; moreover, it can be represented as a summation of two independent versions of Brownian bridges (up to a term of small order of magnitude); see Hsieh and Turnbull (—1996—). In the case of finite samples, $\widehat{ODC}$ does not have the quality we desire for our purpose of estimating the threshold value at a particular operating point.

First, we assume that we have the ODC—but with no knowledge of the individual distributions G and F—and we will propose an empirical procedure or "algorithm" for estimating the threshold at a particular operating point with the aid of $\widehat{ODC}$. This allows us to discover any optimality

properties of the proposed procedure before involving any asymptotic behavior associated with the ODC estimation. Afterwards, we can approximate the ODC by any smoothing method.

CHAPTER 2

Algorithm for Threshold Estimation

2.1. Preliminaries

In this section we propose a procedure (algorithm) that estimates the threshold at a particular operating point M, whose coordinates on the curve are (p, s). Primitive statistical definitions are needed for this purpose; we will introduce them here for the sake of completeness for a wider readership.

The empirical (sample) distribution function G_n, an estimator for G, is defined as

$$G_n(\xi) = \frac{1}{n}\Sigma_i I_{(x_i \leq \xi)},$$

where I is the indicator function. It is known that

$$G_n(\cdot) \overset{a.s.}{\to} G(\cdot),$$

$$n^{1/2}(G_n(\xi) - G(\xi)) \overset{D}{\to} \mathcal{N}(0, G(\xi)(1 - G(\xi))). \tag{2.1}$$

This can be shown to be also true when ξ is a r.v.; see Lemma 4. The p^{th}-quantile $G^{-1}(p)$ is defined as

$$G^{-1}(p) = \inf\{x,\ G(x) \geq p\}.$$

The sample p^{th}-quantile $G_n^{-1}(p)$, an estimator of $G^{-1}(p)$, is the p^{th}-quantile of G_n, i.e.,

$$G_n^{-1}(p) = \inf\{x_i,\ G_n(x_i) \geq p,\ i = 1, \cdots, n\}. \tag{2.2}$$

In the sequel we will be using the notation ξ_p and $\widehat{\xi}_p$ interchangeably for $G^{-1}(p)$ and $G_n^{-1}(p)$ respectively. Analogously to G_n, above, we have

$$G_n^{-1}(\cdot) \overset{a.s.}{\to} G^{-1}(\cdot), \tag{2.3}$$

$$n^{1/2}(G_n^{-1}(p) - G^{-1}(p)) \overset{D}{\to} \mathcal{N}(0, p(1-p)/[G'(\xi_p)]^2), \tag{2.4}$$

if G' exists at ξ_p; a similar result is available if G is only left or right differentiable.

Bahadur (—1966—) derived the almost sure convergence (up to a small order of magnitude) of $\widehat{\xi}_p$ to ξ_p. Later, Ghosh (—1971—) introduced a new elegant proof for Bahadur's result (up to a small order of magnitude in probability, which suffices for many applications including our purpose; see Lemma 2. The introductory material above, along with its derivations, can be found, e.g., in Serfling (1980, Ch. 2).

2.2. Pseudocode and Mathematical Formalization

We assume that we are given a finite sample of the decision values from each of the two classes. These values also provide a representation of the empirical distributions G_n and F_m. We also assume the availability of the ODC, to be replaced ultimately by a parametric or semiparametric estimated smooth curve. Figure 2.1 provides a plot of an ODC and $\widehat{ODC}$; the latter is a plot of $F_m(\xi)$ vs. $G_n(\xi)$ as the threshold ξ is moved throughout its range on the decision axis. When ξ crosses a point $x_i \in G_n$ (or F_m) this produces a horizontal (or vertical) increment on $\widehat{ODC}$. Our task is to estimate the threshold corresponding to a given point $M = (p, s)$ on the ODC.

A natural candidate algorithm for the task of estimating ξ is Algorithm 1. First, find the quantiles in the appropriate distribution that correspond to the upper intersection point, $U = (p_U, s_U)$ and the lower intersection point, $L = (p_L, s_L)$, of the ODC and the $\widehat{ODC}$; a reasonable estimate would then consist of a weighted sum of these quantiles. As we analyze the properties of this estimator below, an alternative candidate estimator will naturally suggest itself. Algorithm 1 is a heuristic starting point, not designed to achieve any optimality property. However, it is a natural startup and an intuitive first-order approximation.

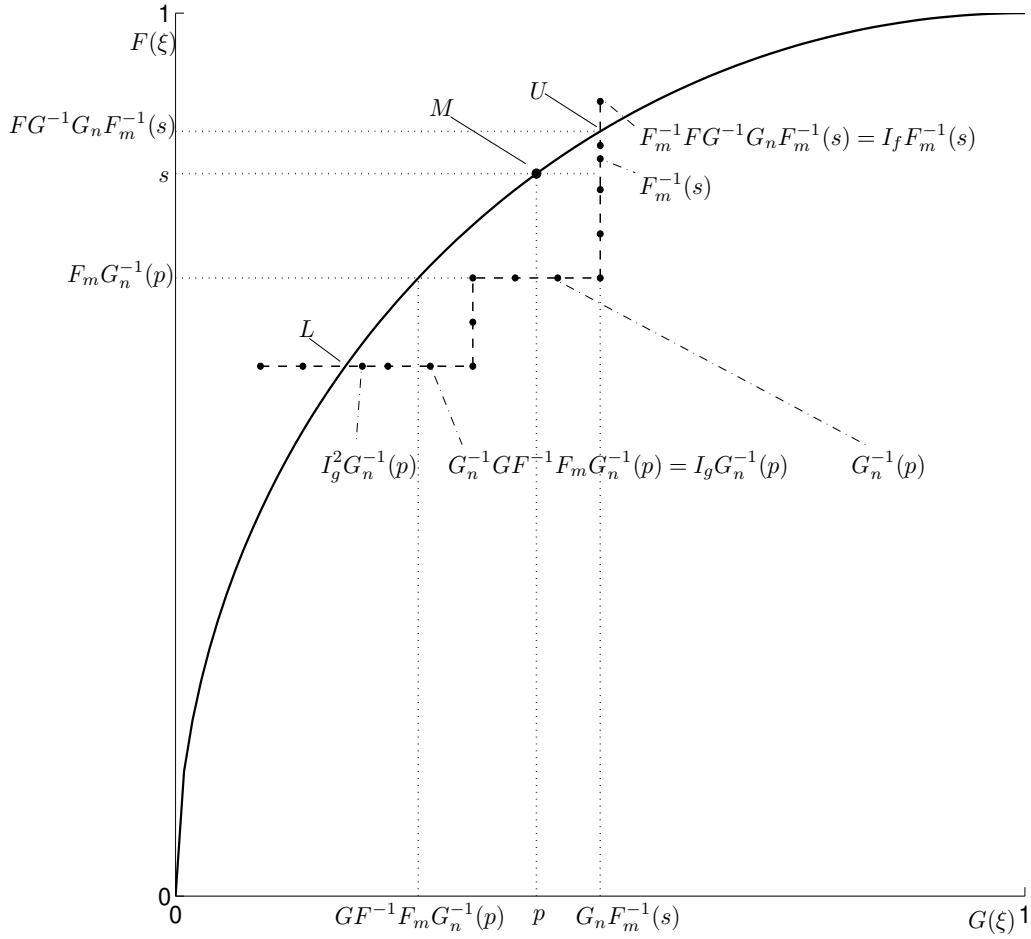


Figure 2.1. The true ODC (solid line) and the empirical one, $\widehat{ODC}$ (staircase line). Starting from an operating point of interest $M = (p, s)$ and moving in the two directions, along the ODC, until intersecting with the $\widehat{ODC}$ gives two estimates of the threshold of interest. These estimates are $\mathrm{I}_g^2\, G_n^{-1}(p)$ and $\mathrm{I}_f\, F_m^{-1}(s)$ in the current case, where $\mathrm{I}_g(\cdot) = G_n^{-1}GF^{-1}F_m^{-1}(\cdot)$ and $\mathrm{I}_f(\cdot) = F_m^{-1}FG^{-1}F_n^{-1}(\cdot)$. The dashed annotations are the quantile values that correspond to the empirical points.

Other ad-hoc procedures are possible. For example, we could have projected every point on the $\widehat{ODC}$ to the ODC and selected the closest two projections from each side (on the curve) of the point M. Then, interpolate on the arc between the threshold values of the two original points (or even the estimated threshold of the two projections). This procedure has the following two problems in the finite sample size. First, there can be more than one projection from a point to the curve; this may happen if the point is a center of a circular arc on the curve. Second, it can happen that two

Algorithm 1 Threshold Estimation at a point $M(=(p,s))$ on the ODC

$\widehat{ODC} \leftarrow$ empirical ODC.
$ODC \leftarrow$ true ODC.
$U(=(p_U, s_U)) \leftarrow$ upper intersection point between ODC and $\widehat{ODC}$.
$L(=(p_L, s_L)) \leftarrow$ lower intersection point between ODC and $\widehat{ODC}$.
if $\widehat{ODC}$ at U is vertical **then**
  $\widehat{\tau}_U = F_m^{-1}(s_U)$ //sample ${s_U}^{th}$-quantile for F.
  $\widehat{\tau}_L = G_n^{-1}(p_L)$ //sample ${p_L}^{th}$-quantile for F.
else
  $\widehat{\tau}_U = G_n^{-1}(p_U)$.
  $\widehat{\tau}_L = F_m^{-1}(s_L)$.
end if
return $\widehat{\xi} = a\widehat{\tau}_U + (1-a)\widehat{\tau}_L$ //A convex combination, which could be obtained, in an ad-hoc way, from interpolation on the curve.

empirical points generate two projections which have reverse order on the ODC, a matter that does not preserve monotonicity of the ODC. We do not study such a procedure in the present work.

A mathematical formalization of Algorithm 1 is straightforward to follow on Figure 2.1; remember that our goal is to give the estimate $\widehat{\xi}$ that estimates ξ $(= G^{-1}(p) = F^{-1}(s))$. The point M $(=(p,s))$ has a sample p^{th}-quantile $G_n^{-1}(p)(=\widehat{\xi_p})$. The empirical distribution F_m evaluated at this threshold is $F_m G_n^{-1}(p)$. There will then be a point on the ODC with the coordinates $GF^{-1}F_m G_n^{-1}(p)$ and $F_m G_n^{-1}(p)$; the corresponding sample quantile at this threshold for ω_0 is then $G_n^{-1}GF^{-1}F_m G_n^{-1}(p) = \mathrm{I}_g\, G_n^{-1}(p)$, where $\mathrm{I}_g(\cdot)$ (an Iteration) is $G_n^{-1}GF^{-1}F_m(\cdot)$. If there is no intersection between the ODC and the $\widehat{ODC}$ at the value $F_m G_n^{-1}(p)$, then we must iterate with the I_g operator starting from the point $\mathrm{I}_g\, G_n^{-1}(p)$ until we find the first intersection. If the first intersection occurs after i iterations, then the estimator $\widehat{\tau}_L$ in Algorithm 1 will be $\mathrm{I}_g^i\, G_n^{-1}(p)$, where I_g^i means applying the operator I_g i-times to its argument. Figure 2.1 shows iterations up to $i=2$. Analogous formulation is immediate for $\widehat{\tau}_U$, with the result of $\mathrm{I}_f^j\, F_n^{-1}(s)$, where $\mathrm{I}_f(\cdot) = F_m^{-1}FG^{-1}G_n(\cdot)$. Note that i and j are random sequences which are functions of the data and m and n. We have not subscripted i and j to simplify the notation. Moreover, we will argue that they will not play a significant role in the analysis in the next section.

CHAPTER 3

Probabilistic Analysis for the Threshold Estimation Algorithm

In this section we provide a rigorous analysis for the asymptotic properties of Algorithm 1; we begin with a roadmap. Since a representative intersection point of ODC and $\widehat{ODC}$ (say L in Figure 2.1) has the estimated threshold of $\mathrm{I}_g^2\, G_n^{-1}(p)$, we can analyze $\mathrm{I}_g\, X_n$ for any arbitrary X_n (a random sequence) and $\mathrm{I}_g^i\, X_n$ can be derived recursively for any i.

We note that the first operation in $\mathrm{I}_g(X_n) = G_n^{-1}GF^{-1}F_m(X_n)$ is $F_m(X_n)$. Thus, we will first require a generalization of (2.1) for $F_m(X_n)$; this is provided by Lemma 4. Analysis of the subsequent operations, GF^{-1}, can then be obtained by a direct application of the delta method. Properties of the final operation, G_n^{-1}, require a generalization of (2.4) to allow for random p; this is provided by Lemma 5. An intermediate generalization of (2.4) limited to a deterministic variable is given in Lemma 2. Having derived the asymptotic behavior for $\mathrm{I}_g^i(G_n^{-1}(p))$ and the dually defined expression $\mathrm{I}_f^j(F_m^{-1}(s))$, then we have the asymptotic behavior of the vector $\widehat{\boldsymbol{\xi}} = (\mathrm{I}_g^i\, G_n^{-1}(p), \mathrm{I}_f^j\, F_m^{-1}(s))'$. The two components of $\widehat{\boldsymbol{\xi}}$ will be shown to be asymptotically unbiased for ξ; hence $\widehat{\boldsymbol{\xi}}$ is unbiased for $\xi\mathbf{1}$, where $\mathbf{1}$ is a vector of two components, each equals one. We can then solve for the linear combination $\boldsymbol{\alpha}$ that minimizes the asymptotic variance of the proposed estimator $\boldsymbol{\alpha}'\widehat{\boldsymbol{\xi}}$, which will be asymptotically unbiased for ξ as well.

***Theorem* 1.** Consider the quantile process $\widehat{\boldsymbol{\xi}} = (\mathrm{I}_g^i\, G_n^{-1}(p), \mathrm{I}_f^j\, F_m^{-1}(s))'$ produced by Algorithm 1; then

(1) the estimator $\widehat{\boldsymbol{\xi}}$ has the following weak convergence property, as $n \to \infty$, $m \to \infty$

$$n^{1/2}(\widehat{\boldsymbol{\xi}} - \xi\mathbf{1}) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), where$$

$$\Sigma = \begin{pmatrix} (i+1)^2 e_g^2 + i^2 \frac{e_f^2}{\lambda} & -(i+1)je_g^2 - i(j+1)\frac{e_f^2}{\lambda} \\ -(i+1)je_g^2 - i(j+1)\frac{e_f^2}{\lambda} & (j+1)^2 \frac{e_f^2}{\lambda} + j^2 e_g^2 \end{pmatrix},$$

$$e_g{}^2 = p(1-p)/G'^2(\xi),$$

$$e_f{}^2 = s(1-s)/F'^2(\xi),$$

and we assume that m/n converges, say to λ as $n \to \infty$.

(2) For every i and j there is a linear combination $\alpha'(i,j)$ that minimizes the asymptotic variance of $\boldsymbol{\alpha}'\widehat{\boldsymbol{\xi}}$; and this minimum variance does not depend on either i or j and is given by

$$\min_{\boldsymbol{\alpha}} \text{var}[\boldsymbol{\alpha}'\widehat{\boldsymbol{\xi}}] = \frac{e_g^2 e_f^2}{\lambda e_g^2 + e_f^2};$$

The value of $\boldsymbol{\alpha}$ that achieves this minimum variance can be obtained directly (without iterations) from the case $i = j = 0$ (the case of sample quantiles $(G_n^{-1}(p), F_m^{-1}(s))$, as

$$\boldsymbol{\alpha} = \left(\frac{1}{1 + \frac{e_g^2}{e_f^2/\lambda}}, \frac{1}{1 + \frac{e_f^2/\lambda}{e_g^2}} \right)'. \tag{3.1}$$

Proof. Consider the point $M = (p, s)$ on ODC with a corresponding threshold ξ, where $F(\xi) = s$, and $G(\xi) = p$. Then, for any X_n such that $n^{1/2}(X_n - \xi) \xrightarrow{D} X$, by applying Lemma 4 with F_m in place of G_n of the Lemma, we get

$$m^{1/2}(F_m(X_n) - s) = m^{1/2}(F_m(\xi) - s) + m^{1/2}F'(\xi)(X_n - \xi) + o_p(1). \tag{3.2}$$

Then a direct application of the delta method to the function GF^{-1}, whose derivative is G'/F', and normalizing the rate of convergence to $n^{1/2}$, leads to

$$n^{1/2}(GF^{-1}F_m(X_n) - p) = \frac{m^{1/2}}{\lambda^{1/2}} \frac{G'(\xi)}{F'(\xi)} (F_m(\xi) - s) + n^{1/2}G'(\xi)(X_n - \xi) + o_p(1). \tag{3.3}$$

Now, by direct application of Lemma 5 with $p_n = GF^{-1}F_m(X_n)$ we get

$$n^{1/2}(G_n^{-1}GF^{-1}F_m(X_n) - \xi) = \frac{m^{1/2}}{\lambda^{1/2}F'(\xi)}(F_m(\xi) - s) + n^{1/2}(X_n - \xi) - \frac{n^{1/2}(G_n(\xi) - p)}{G'(\xi)} + o_p(1)$$
$$= -e_g N_g + e_f/\sqrt{\lambda}N_f + X + o_p(1), \tag{3.4}$$

where $n^{1/2}(G_n(\xi) - p) \xrightarrow{D} \mathcal{N}(0, p(1-p))$; and N_g is its normalized version. Similar definitions hold for N_f.

If we denote $G_n^{-1}GF^{-1}F_m(\cdot)$ by the operator $\mathrm{I}_g(\cdot)$ then it is obvious that multifolded application—in the language of computer science is called recursion—of I_g i-times to X_n, i.e., $\mathrm{I}_g^i(X_n)$, gives the following property

$$n^{1/2}(\mathrm{I}_g^i(X_n) - \xi) = -ie_g N_g + ie_f/\sqrt{\lambda}N_f + X + o_p(1). \tag{3.5}$$

Now if we choose X_n to be the sample quantile $G_n^{-1}(p)$, then $X = -\sqrt{e_g}N_g$ by Lemma 2; and (3.5) can be rewritten as

$$n^{1/2}(\mathrm{I}_g^i\, G_n^{-1}(p) - \xi) = -(i+1)e_g N_g + ie_f/\sqrt{\lambda}N_f + o_p(1). \tag{3.6}$$

The following analogous relation is immediate,

$$n^{1/2}(\mathrm{I}_f^j\, F_m^{-1}(s) - \xi) = je_g N_g - (j+1)e_f/\sqrt{\lambda}N_f + o_p(1). \tag{3.7}$$

Since N_g and N_f are independent, then part (1) is proved. We should note that with higher iterations on the ODC (higher values of i and j) the variance of each component of $\widehat{\boldsymbol{\xi}}$ increases, however the covariance decreases (negative covariance). Part (2) is proved by direct application of Lemma 6. Notice that $\boldsymbol{\alpha}$ does not depend explicitly on G' or F', rather, it depends on the ratio $e_g^2/e_f^2 = p(1-p)/(s(1-s)S^2(p,s))$, where $S(p,s)$ is the slope of the ODC at the point (p,s), which is assumed to be known. Notice that e_g^2 and e_f^2 are the asymptotic variances of $G_n^{-1}(p)$ and $F_m^{-1}(s)$ respectively. ∎

This theorem indicates that the proposed algorithm has the same asymptotic efficiency as simply using the sample quantiles from the respective distributions at the point (p,s) with the convex

combination given in (3.1). Moreover, the details of the proofs of the Lemmas (see Appendix) that are used in the proof of the above theorem reveal that the o_p term in (3.6) is an accumulation of those in (3.2)–(3.4). Similarly is the o_p term in (3.7). A corresponding theorem for the case of finite samples in this problem does not seem tractable. However, qualitatively speaking, one can anticipate that the asymptotically vanishing terms will accumulate into variances in the finite sample problem. This means that it is anticipated to be more efficient to stop with the sample quantiles $(G_n^{-1}(p), F_m^{-1}(s))$, i.e., the case where $i = j = 0$, together with the linear combination of (3.1). In terms of the steps of Algorithm 1, the best performance will result without taking any iterations, i.e., with considering only the estimates $G_n^{-1}(p)$ and $F_m^{-1}(s)$. In Section 4 we provide results of simulations that support this anticipated behavior.

One can improve upon the above estimator by using a better quantile estimator than the simple sample quantile (2.2). If we use unbiased quantile estimators $(Q_n^{-1}(p), Q_m^{-1}(s))$, whose asymptotic variances are $(\sigma_g^2, \sigma_f^2/\lambda)$ with zero covariance, in place of $(G_n^{-1}(p), F_m^{-1}(s))$, the best linear combination would be (3.1) with σ_g^2 and σ_f^2 in place of e_g^2 and e_f^2 respectively. In this case $\sigma_g^2 = \sigma_g^2(p, G)$, and similarly σ_f^2, in which we do not guarantee that σ_g^2/σ_f^2 will be only a function of the ODC; it may require knowledge of the individual distributions G and F. However, a good estimator for the variance is available via bootstrapping. If we still use $\boldsymbol{\alpha}$ in (3.1) with the new quantile estimator it is obvious that the linear combination will still remain unbiased with less variance since $e_g < \sigma_g$ and $e_f < \sigma_f$ (even though not the best). The majority of the techniques in the literature for deriving an estimator better than the sample quantile utilize the information available from the order statistics—$G_n^{-1}(p)$ is only one of them—and try to find a linear combination of them; the estimate will be a kernel-based weighted sum of the order statistics close to $G_n^{-1}(p)$ (see Harrell and Davis, —1982—; Falk, —1985—).

All of the analysis above assumes the knowledge of the true ODC, however, in practice it is difficult to realize this condition. There are many techniques available in the literature that fit the ROC curve (the analogue to the ODC). Among them are, Metz (—1986—) (an implementation of Dorfman and Alf (—1969—)) and its newer version, Metz and Pan (—1999—), that guarantees

convexity. These two fitting techniques assume that the two distributions can be transformed simultaneously to two normal distributions. Qin and Zhang (—2003—) assumes the ratio between the two probability densities to be exponential; this allows the use of the logistic regression. Pepe (—2000—) suggests direct fitting to the empirical ROC curve by using the Generalized Linear Models (GLM). Other nonparametric methods are available as well (c.f. Lloyd and Yong, —1999—).

For every p there is a corresponding $\widehat{s}$ from the fitted ODC, and vice versa, for every s there is a corresponding $\widehat{p}$. In terms of just one of them (let's say $\widehat{s}$) we have $\widehat{s} = \widehat{s}(p, F_m, G_n)$; and consequently $F_m^{-1}(\widehat{s})$ is no longer independent of $G_n^{-1}(p)$. We should notice that the asymptotic (and the finite sample size) performance of $\widehat{s}$ is an intrinsic property of the fitting algorithm. In general, assume that the fitting algorithm yields the weak convergence of $\widehat{s}$ such that

$$n^{1/2}(\widehat{s} - s) \xrightarrow{D} S,$$

where $S = S(p, F, G)$. Then direct application of Lemma 4 shows that

$$n^{1/2}(F_m^{-1}(\widehat{s}) - F^{-1}(s)) \xrightarrow{D} \frac{S}{F'(F^{-1}(s))} - \sqrt{e_f/\lambda} N_f + o_p(1). \tag{3.8}$$

This means that if the fitting algorithm is asymptotically unbiased, i.e., S has a zero mean, then $F_m^{-1}(\widehat{s})$ will be asymptotically unbiased for $F^{-1}(s)$ $(= \xi)$. It should be obvious from (3.8) that, the estimator $F_m^{-1}(\widehat{s})$ has an asymptotic variance that is inversely proportional to $F'(\xi)$. This means we will get the most out of it when ξ is in a neighborhood of a high density f. The best linear combination for $G_n^{-1}(s)$ and $F_m^{-1}(\widehat{s})$ depends on S, i.e., the fitting algorithm, and its asymptotic variance and covariances with both N_g and N_f. Moreover, for the finite sample size problem this combination may not work well since it is derived from the asymptotic distribution. Rather, we will rely in our simulations on estimating the covariance structure, and hence the best linear combination, between $G_n^{-1}(s)$ and $F_m^{-1}(\widehat{s})$ by bootstrapping. Then the estimator should be in the form:

$$\widehat{\boldsymbol{\alpha}}_b'(G_n^{-1}(p), F_m^{-1}(\widehat{s})),$$

where $\widehat{\boldsymbol{\alpha}}_b'$ is the bootstrap estimate of the best linear combination given by Lemma 6 for the two estimators $G_n^{-1}(p)$ and $F_m^{-1}(\widehat{s})$. If the data is censored such that the largest order statistic $x_{(m)}$

from the distribution F is smaller than the estimator $G_n^{-1}(p)$—or, vise versa, if the smallest order statistic $y_{(1)}$ from the distribution G is larger than $F_m^{-1}(s)$ in case of estimating the threshold that corresponds to a particular s—we should expect the estimator to be biased. Even if the current two samples are not censored, bootstrapping can reveal whether this will be the case or not if the majority (say 90%) of the bootstrap replications are not censored; we call this condition $cond_B$. It is easy to almost eliminate this bias by proposing the final version of our estimator as:

$$\widehat{\xi}(p) = \begin{cases} \widehat{\boldsymbol{\alpha}}_b'(G_n^{-1}(p), F_m^{-1}(\widehat{s})) & if \quad G_n^{-1}(p) < x_{(m)} \ \& \ cond_B \\ G_n^{-1}(p) & otherwise \end{cases}, \tag{3.9a}$$

$$\widehat{\xi}(s) = \begin{cases} \widehat{\boldsymbol{\alpha}}_b'(F_m^{-1}(s), G_n^{-1}(\widehat{p})) & if \quad y_{(1)} < F_m^{-1}(s) \ \& \ cond_B \\ F_m^{-1}(s) & otherwise \end{cases}. \tag{3.9b}$$

It should be noted that, by estimating the ROC we are no longer able to define an operating point on the curve with full determinism, i.e., by specifying a pair of (FPF, TPF), which equals to (1-s, 1-p). Rather, we can define one of these two fractions and get an estimate of the other from the fitted ROC. In that sense, the problem of the threshold estimation can be viewed as quantile estimation for the specified fraction. The utility of the other distribution, say F, is then to account as a smoothing kernel that is used to "learn" the G distribution. If the two curves are separated enough we do not get an advantage of F as mentioned. This is obvious from (3.9), since F' would be small and increases the variance of $F_m^{-1}(\widehat{s})$. We cannot behave worse than $G_n^{-1}(p)$ as the linear combination is convex; and this is provided that the fitting procedure is unbiased as declared. If the fitting algorithm is biased for some range of p, as will be demonstrated in Section 4, the sample quantile may outperform our estimator.

CHAPTER 4

Simulation

In this section we present the simulation results carried out to assess the proposed estimator. We utilized different distributions to span a wide range of possibilities. Table 4.1 lists those experiments and the parameters of the corresponding distributions. The F and G distributions in experiment 1 (Exp. 1) are normals with common variance equals 1 and means equal 0 and 1 respectively. This separation between the two means achieves around 0.76 AUC (Area Under the Curve). This is the area under the ROC or ODC, since $\int TPF \; dFPF = \int s \; dp$. The other experiments utilize the "Generalized Lambda Distribution" (Ramberg et al., —1979—). This distribution is characterized by 4 different parameters, namely, $\lambda_1, \cdots, \lambda_4$; they determine the shape of the distribution. The four columns in the table display those lambda values for the two distributions. The two distributions in Exp. 2 are skewed to the left and to the right respectively (skewed outside), in Exp. 3 are skewed to the right, and in Exp. 4 are skewed to the left. In Exp. 5 the two distributions are symmetric with heavy tails similar to those of the t-dist with 9 df. The scale parameter λ_1 of F is always set to 0 and that of G is adjusted to achieve the separability that gives the same value (0.76) of the AUC of Exp. 1.

In section 3 we saw that the estimator obtained using Algorithm 1 with iterations on the true ODC is asymptotically as efficient as without iterations. In that section we anticipated that the finite sample size performance without iterations will be even better. Figure 4.1 illustrates this fact by plotting the relative efficiency (RE) for one- vs. no-iteration estimators. This efficiency is plotted vs. the TPF, i.e., $(1-p)$, for different values of the sample size $n = m$. The results are obtained via 10,000 trials of MC study for Exp. 1.

The RE between the estimator (3.9) and the sample quantile $G_n^{-1}(p)$ is illustrated in Figure 4.2. The number of MC trials is 10,000. For every MC trial, one sample from every distribution, F and G, is simulated and the ROC is fitted using the GLM, where both the link and basis functions used are the *probit*. Then, the two estimators in (3.9), $G_n^{-1}(p)$ and $F_m^{-1}(\widehat{s})$, are obtained and 200 bootstrap replications are used to get the estimate $\widehat{\alpha}_b$. This procedure gives one MC estimate of the estimator (3.9).

The results remarkably illustrate the superiority of the proposed estimator over the sample quantile, especially for the high values of the TPF. In Exp. 2 the RE drops below 1 for larger sample size n, in some range of the TPF. This is due to the fact that the major part of the ODC, under this configuration, looks like a straight line with slope close to 1. This vanishing curvature is challenging for the undertaken fitting procedure to cope with. This produces bias except where the true and the fitted ROC intersect; this violates the assumption of having an unbiased fitting procedure; hence the bias of the estimator and the large MSE. This pattern appears again in Figure 4.2.e, though less severely.

The results illustrated in these figures are produced, as aforementioned, under distribution separation that resulted in an AUC value of 0.76. From the discussion in the previous section and above it is expected that the efficiency of the estimator will be higher with less separated distributions, i.e., with lower AUC. Figure 4.3 illustrates the efficiency of the estimator for Exp. 1, when the sample size $n = 30$. The efficiency is plotted vs. the TPF under three different values of AUC, 0.60, 0.76, and 0.90, by increasing the mean of G. The lower the AUC the more efficient the estimator. This observation is common to all other values of the sample size n, and to all other experiments as well.

Exp.	F, G Parameters				Skewness
1	$\mathcal{N}(0,1), \mathcal{N}(1,1)$				symmetric
2	0,-.035	.034	.0285,.009695	.009695,.0285	outside
3	0,.91	.034	.009695	.0285	right
4	0,.91	.034	.0285	.009695	left
5	0,.86	-.3203	-.1359	-.1359	symmetric heavy tails

Table 4.1. Description of different experiments.The Generalized Lambda Distribution is used in exp. 2–5; e.g., Exp. 2 has λ_1 equals 0 and $-.035$ for F and G respectively; while, for both distributions, $\lambda_2 = .034$. λ_3 and λ_4 are displayed in the next two columns. The skewness of the two distributions is indicated in the last column; e.g., outside skewness indicates that F is left-skewed and G is right-skewed. Exp. 5 utilizes heavy-tailed distributions similar to t-dist with 9 df.

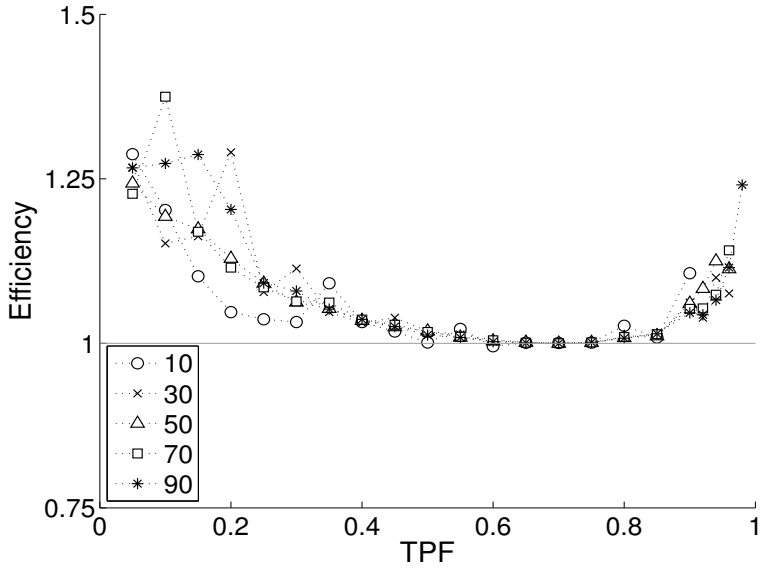
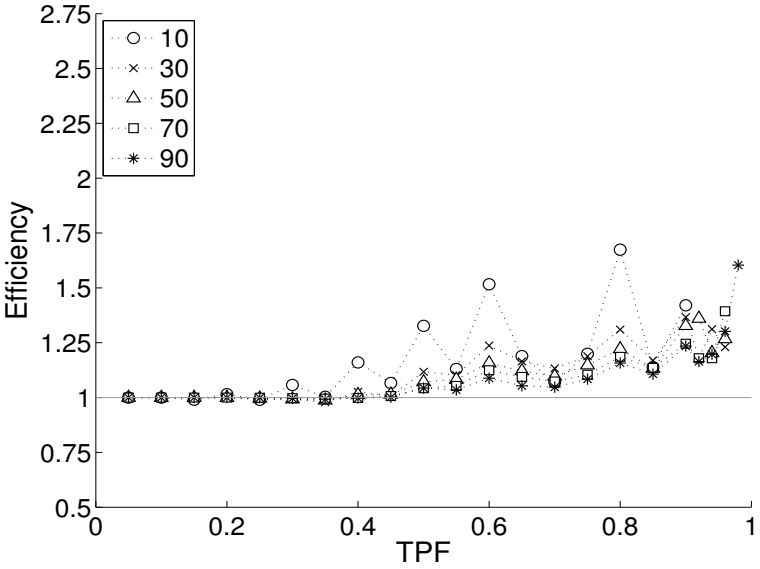
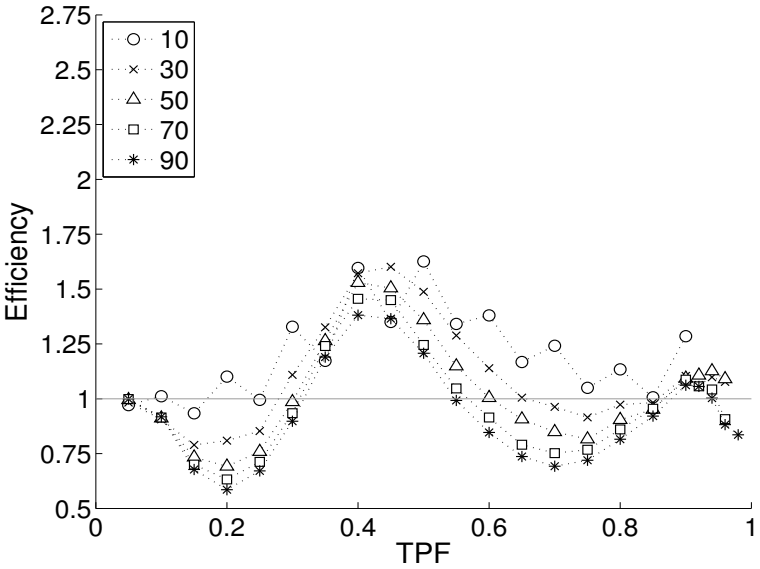


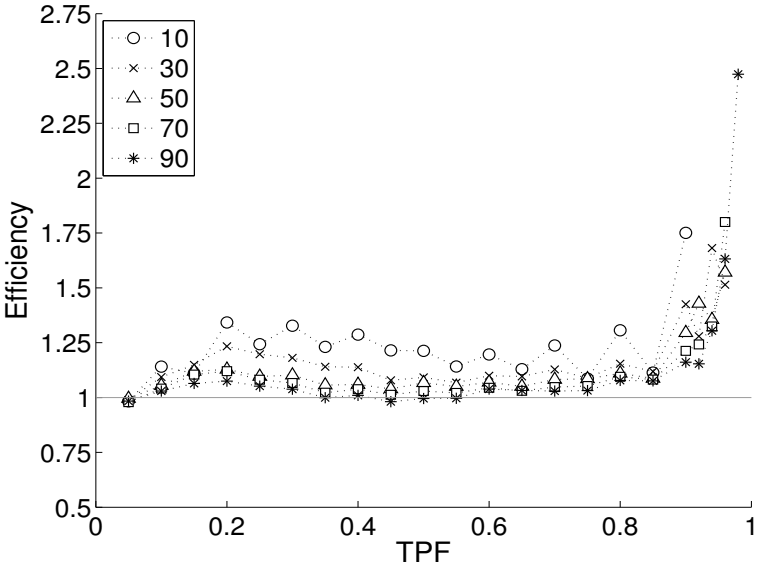
Figure 4.1. The relative efficiency between two estimators for exp. 1 assuming the knowledge of the true ODC. The first estimator does not iterate on the true ODC; the second iterates once. The legend indicates five values of the sample size $n = m$


10
30
50
70
90
Efficiency
TPF
2.75
2.5
2.25
2
1.75
1.5
1.25
1
0.75
0.5
0
0.2
0.4
0.6
0.8
1

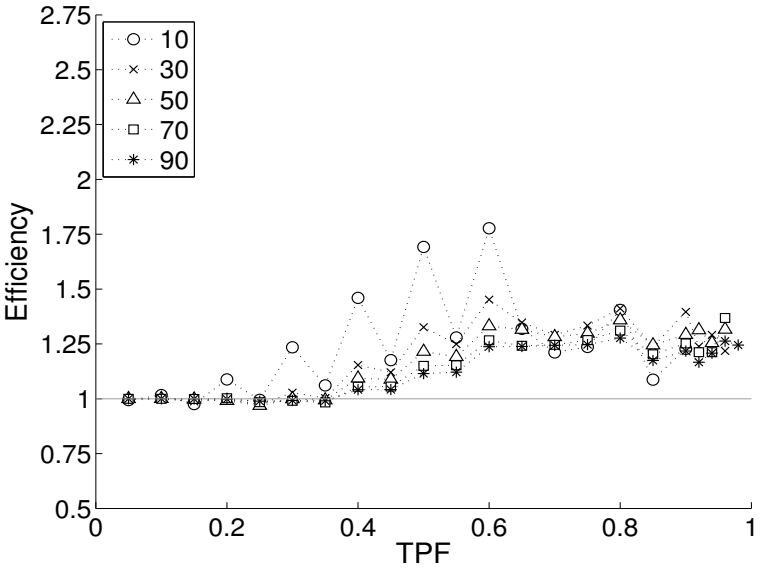
(4.2.a)


10
30
50
70
90
Efficiency
TPF
2.75
2.5
2.25
2
1.75
1.5
1.25
1
0.75
0.5
0
0.2
0.4
0.6
0.8
1

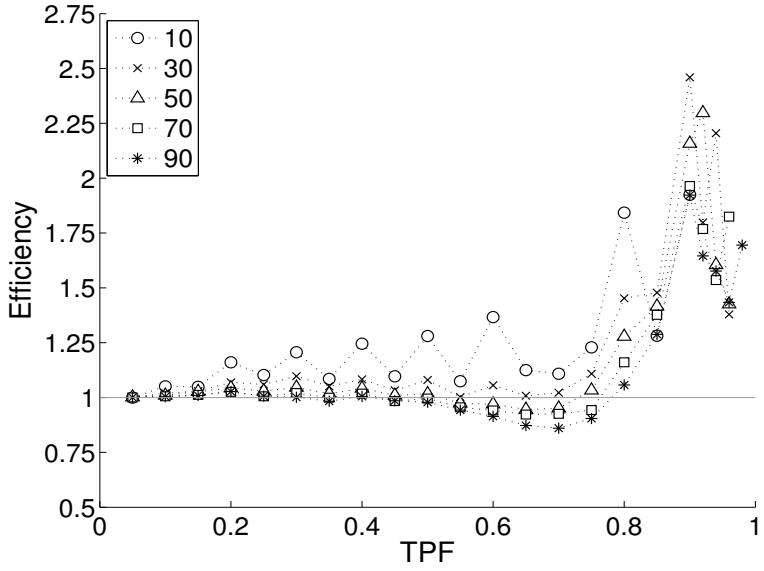
(4.2.b)


10
30
50
70
90
Efficiency
TPF
2.75
2.5
2.25
2
1.75
1.5
1.25
1
0.75
0.5
0
0.2
0.4
0.6
0.8
1

(4.2.c)


10
30
50
70
90
Efficiency
TPF
2.75
2.5
2.25
2
1.75
1.5
1.25
1
0.75
0.5
0
0.2
0.4
0.6
0.8
1

(4.2.d)



(4.2.e)

Figure 4.2. The relative efficiency between the proposed estimator and the sample quantile $G_n^{-1}(p)$ for different experiments. The proposed estimator outperforms the sample quantile especially in the high TPF range. The legend indicates five values of the sample size n, where m is kept fixed at 100

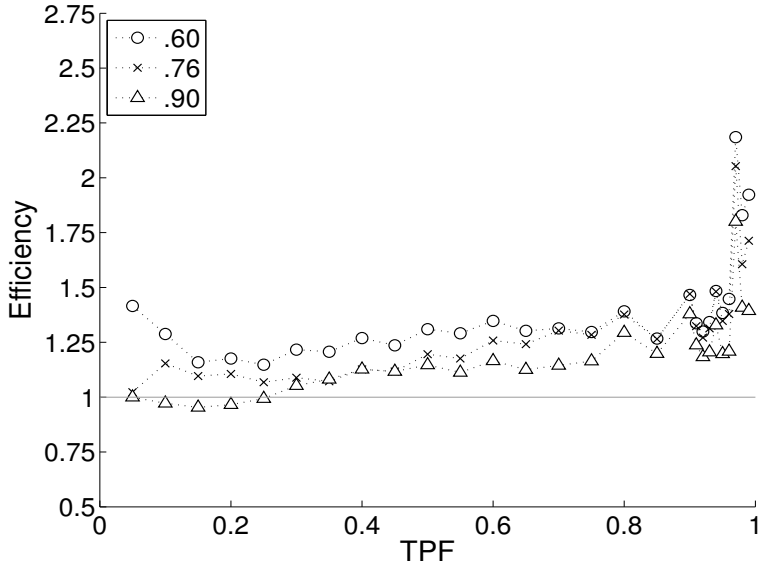


Figure 4.3. Estimator efficiency for Exp. 1 under sample size $n = 30$, for 3 different values of AUC (shown in legend). The smaller the AUC the more efficient the estimator.

CHAPTER 5

Conclusion

The problem of estimating the threshold of a particular operating point on the ODC was considered. At the beginning, a heuristic procedure that utilizes the intersection between the empirical and the true ODC was proposed; the mathematical analysis shows that it is asymptotically as efficient as using only the operating point on the ODC. In practice we can, at best, approximate the true ODC by a fitting procedure; we exploited the GLM for that purpose. The two sample quantiles at an operating point on the fitted curve are no longer independent, and the best linear combination is obtained by bootstrapping the current sample. Simulation results, under different distribution configurations, show the superiority of the proposed estimator over the sample quantile from one distribution.

Appendix: Lemmas and Proofs.

The reader may refer to Sections 1 and 2 for the definitions of all the variables and quantities used in the current section.

***Lemma* 2** (*Ghosh —1971—*)**.** Suppose $G'(\xi_p)$ exists and is strictly positive and $p_n = p + O(n^{-1/2})$, then we have the following property for the sample p_n-quantile

$$G_n^{-1}(p_n) = M_{p_n} - \frac{(G_n(\xi_p) - p)}{G'(\xi_p)} + o_p(n^{-1/2}), \tag{5.1}$$

$$M_{p_n} = \xi_p + \frac{p_n - p}{G'(\xi_p)}. \tag{5.2}$$

If G has a continuous first derivative g in a neighborhood around ξ_p then the Lemma is true for the more relaxed restriction, $p_n = p + o(1)$ and with M_{p_n} being replaced by ξ_{p_n} $(= G^{-1}(p_n))$.

***Lemma* 3.** Assume that there exist two random sequences $V_n(\omega)$ and $W_n(\omega)$, where $W_n(\omega)$ converges weakly, defined on the same probability space, where

$$\{\omega : V_n \leq t\} \equiv \{\omega : W_n + o_p(1) \leq t\}, \forall\ t.$$

Then

$$V_n = W_n + o_p(1).$$

Proof. For all $\varepsilon > 0$, the set $\{V_n \leq x,\ W_n \geq x + \varepsilon\} \equiv \{W_n + R_n \leq x,\ W_n \geq x + \varepsilon\} \subseteq \{R_n \leq -\varepsilon\}$; the latter has a probability tending to zero, as $n \to \infty$, by the definition of $o_p(1)$. Hence

$$\lim_{n\to\infty} \Pr\{V_n \leq x,\ W_n \geq x + \varepsilon\} = 0, \qquad \forall\ \varepsilon > 0. \tag{5.3}$$

A similar argument leads to

$$\lim_{n\to\infty}\Pr\{V_n \geq x+\varepsilon,\ W_n \leq x\} = 0\ \forall\ \varepsilon > 0. \tag{5.4}$$

The conditions (5.3) and (5.4), along with that W_n converges weakly—from the convergence of G_n and from the assumption for X_n—establish the conditions of Lemma 1 of Ghosh (—1971—). Hence, $V_n - W_n \xrightarrow{p} 0$ as was to be proved.

The last argument can be replaced by the following simpler one. By noticing that $W_n + R_n \xrightarrow{D} W$ (by Slutsky's theorem), then we have

$$\begin{aligned}\lim_{n\to\infty}\Pr\{V_n \leq x\} &= \lim_{n\to\infty}\Pr\{W_n + R_n \leq x\}\\ &= \Pr\{W \leq x\}.\end{aligned}$$

Therefore, $V_n \xrightarrow{D} W$,and hence $V_n = W + o_p(1)$. But $W_n = W + o_p(1)$, hence $V_n = W_n + o_p(1)$.

We can see it in a third way by the following. Since the set equality is for every t, then it is obvious that $\{\omega : V_n \in \mathcal{B}\} \equiv \{\omega : W_n \in \mathcal{B}\}\ \forall$ Borel set $\mathcal{B}$. Then $V_n \overset{a.s.}{=} W_n$. ∎

Lemma 4. Suppose $G'(\xi_p)$ exists and is strictly positive and the random sequence X_n has the property that $n^{1/2}(X_n - \xi_p)$ converges weakly, then

$$G_n(X_n) = G_n(\xi_p) + G'(\xi_p)(X_n - \xi_p) + o_p(n^{-1/2}).$$

Proof. The lemma is quite easy to prove, but we include the proof for the sake of completeness.

$$\begin{aligned}\{\omega : n^{1/2}(G_n(X_n) - p) \leq x\} &\equiv \{\omega : X_n \leq G_n^{-1}(p + xn^{-1/2})\}\\ &\equiv \{\omega : n^{1/2}(G_n(\xi_p) - p) + n^{1/2}G'(\xi_p)(X_n - \xi_p) + o_p(1) \leq x\}.\end{aligned}$$

For the first set equivalence, see Serfling (1980, Sec. 1.1.4); the second is a direct substitution from Lemma 2 and arrangements. Hence, by Lemma 3,

$$n^{1/2}(G_n(X_n) - p) = n^{1/2}(G_n(\xi_p) - p) + n^{1/2}G'(\xi_p)(X_n - \xi_p) + o_p(1),$$

and the Lemma follows directly from here. ■

A special case of the above Lemma, when X_n is the median of another distribution, was first derived in Gastwirth (—1968—) using a different technique. We prefer the above method since it is more general and does not need extra assumptions.

***Lemma* 5** (*Generalization to Lemma 2*)**.** Suppose $G'(\xi_p)$ exists and is strictly positive and $n^{1/2}(p_n - p)$ converges weakly, then we have an exactly similar expansion for the sample p_n-quantile as in Lemma 2, i.e.,

$$G_n^{-1}(p_n) = M_{p_n} - \frac{(G_n(\xi_p) - p)}{G'(\xi_p)} + o_p(n^{-1/2}),$$
$$M_{p_n} = \xi_p + \frac{p_n - p}{G'(\xi_p)}.$$

with M_{p_n} is a r.v. rather than deterministic.

Proof. Since

$$\{\omega : n^{1/2}(G_n^{-1}(p_n) - \xi_p) \le t\} \equiv \{\omega : p_n \le G_n(\xi_p + tn^{-1/2})\}$$
$$\equiv \{\omega : n^{1/2}(p_n - p) \le n^{1/2}(G_n(\xi_p + tn^{-1/2}) - p)\},$$

then direct substitution from Lemma 4 with $X_n = \xi_p + tn^{-1/2}$ (a degenerate r.v.) gives

$$\{\omega : n^{1/2}(G_n^{-1}(p_n) - \xi_p) \le t\} \equiv \{\omega : \frac{n^{1/2}(p_n - p)}{G'(\xi_p)} - \frac{n^{1/2}(G_n(\xi_p) - p)}{G'(\xi_p)} + o_p(1) \le t\}.$$

Then

$$n^{1/2}(G_n^{-1}(p_n) - \xi_p) = \frac{n^{1/2}(p_n - p)}{G'(\xi_p)} - \frac{n^{1/2}(G_n(\xi_p) - p)}{G'(\xi_p)} + o_p(1),$$

as was to be proved ■

***Lemma* 6** (*Best Unbiased Linear Combination*)**.** If $\widehat{\boldsymbol{\xi}}$ is unbiased for $\xi\mathbf{1}$, with a covariance structure Σ, then the best unbiased linear combination (under square loss) $\boldsymbol{\alpha}'\widehat{\boldsymbol{\xi}}$ has the minimum variance of $1/(\mathbf{1}'\Sigma^{-1}\mathbf{1})$ when $\boldsymbol{\alpha} = (\mathbf{1}'\Sigma^{-1})/(\mathbf{1}'\Sigma^{-1}\mathbf{1})$.

Proof. The lemma is quite easy to prove. To have $\boldsymbol{\alpha}'\widehat{\boldsymbol{\xi}}$ unbiased $\forall\ \boldsymbol{\xi}$ is equivalent to the constraint $\boldsymbol{\alpha}'\mathbf{1}$ =1. The variance σ^2 of the linear combination $\boldsymbol{\alpha}'\widehat{\boldsymbol{\xi}}$ can be written with Lagrange multiplier of that constraint as

$$\sigma^2 = \boldsymbol{\alpha}'\Sigma\boldsymbol{\alpha} + \lambda(\boldsymbol{\alpha}'\mathbf{1}-1).$$

Direct differentiation and substitution gives the result. ■

References

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*, Wiley series in probability and statistics, 3rd edn, Wiley-Interscience, Hoboken, N.J.

Bahadur, R. R. (—1966—). A note on quantiles in large samples, *The Annals of Mathematical Statistics* **37**(3): 577–580. 00034851 Institute of Mathematical Statistics.

Dorfman, D. D. and Alf, E. (—1969—). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals–rating method data., *Journal of Mathematical Psychology* **6**: 487–496.

Falk, M. (—1985—). Asymptotic normality of the kernel quantile estimator, *The Annals of Statistics* **13**(1): 428–433. Short Communications 00905364 Institute of Mathematical Statistics.

Gastwirth, J. L. (—1968—). The first-median test: A two-sided version of the control median test, *Journal of the American Statistical Association* **63**(322): 692–706. 01621459 American Statistical Association latex.

Ghosh, J. K. (—1971—). A new proof of the bahadur representation of quantiles and an application, *The Annals of Mathematical Statistics* **42**(6): 1957–1961. 00034851 Institute of Mathematical Statistics latex.

Harrell, F. E. and Davis, C. E. (—1982—). A new distribution-free quantile estimator, *Biometrika* **69**(3): 635–640. 00063444 Biometrika Trust latex.

Hsieh, F. and Turnbull, B. W. (—1996—). Nonparametric and semiparametric estimation of the receiver operating characteristic curve, *Annals of Statistics* **24**(1): 25–40.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*, Springer texts in statistics, 3rd edn, Springer, New York.

Lloyd, C. J. and Yong, Z. (—1999—). Kernel estimators of the ROC curve are better than empirical, *Statistics and Probability Letters* **44**(3): 221.

Metz, C. E. (—1986—). Statistical analysis of ROC data in evaluating diagnostic performance, *in* D. E. Herbert and R. H. Myers (eds), *Multiple regression analysis: Applications in the health sciences.*, New York, pp. 365–384.

Metz, C. E. and Pan, X. (—1999—). "proper" binormal ROC curves: Theory and maximum-likelihood estimation., *Journal of Mathematical Psychology* **1**: 1–33.

Pepe, M. S. (—2000—). An interpretation for the ROC curve and inference using glm procedures, *Biometrics* **56**(2): 352–359.

Qin, J. and Zhang, B. (—2003—). Using logistic regression procedures for estimating receiver operating characteristic curves, *Biometrika* **90**(3): 585–596.

Ramberg, J. S., Tadikamalla, P. R., Dudewicz, E. J. and Mykytka, E. F. (—1979—). A probability distribution and its uses in fitting data, *Technometrics* **21**(2): 201. 0040-1706 Article type: Full Length Article / Full publication date: May, 1979 (197905). / Copyright 1979 American Society for Quality and American Statistical Association.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*, Wiley series in probability and mathematical statistics, Wiley, New York.