Kauppi, Heikki

Working Paper

The Generalized Receiver Operating Characteristic Curve

Discussion paper, No. 114

Provided in Cooperation with:
Aboa Centre for Economics (ACE), Turku

Suggested Citation: Kauppi, Heikki (2016) : The Generalized Receiver Operating Characteristic Curve, Discussion paper, No. 114, Aboa Centre for Economics (ACE), Turku

This Version is available at:
https://hdl.handle.net/10419/233329

Heikki Kauppi

The Generalized Receiver Operating Characteristic Curve

Aboa Centre for Economics

Discussion paper No. 114

Turku 2016

The Aboa Centre for Economics is a joint initiative of the economics departments of the University of Turku and Åbo Akademi University.

ISSN 1796-3133

Printed in Uniprint
Turku 2016

Heikki Kauppi

The Generalized Receiver Operating Characteristic Curve

Aboa Centre for Economics
Discussion paper No. 114
October 2016

ABSTRACT

The problem is to predict whether a random outcome is a "success" ($R = 1$) or a "failure" ($R = 0$) given a continuous variable Z. The performance of a prediction rule $D = D(Z) \in \{1, 0\}$ boils down to two probabilities, $\beta = \Pr(D = 1|R = 1)$ and $\alpha = \Pr(D = 1|R = 0)$. We wish β is high, α is low. Given a set of rules $\mathcal{D}$ such that any $D \in \mathcal{D}$ is attributed to a specific α, I define the "generalized" receiver operating characteristic (GROC) curve as a function that returns β for any $\alpha \in [0, 1]$. The GROC curve associated with $\mathcal{D} = \{D(Z) = I(Z > c), c \in \mathbb{R}\}$ is the "conventional" ROC curve, while an "efficient" ROC (EROC) curve derives from rules that return the largest possible β for any $\alpha \in [0, 1]$. I present estimation theory for the GROC curve and develop procedures for estimating the efficient rules and the associated EROC curve under semiparametric and nonparametric conditions.

Keywords: classification problem, receiver operating characteristic (ROC) curve, likelihood ratio rule, semi-parametric estimation, non-parametric estimation

Contact information

Heikki Kauppi
Department of Economics
University of Turku
FI-20014, Finland
Email: heikki.kauppi (at) utu.fi

Acknowledgements

I thank Hannu Oja and Pentti Saikkonen for useful comments on earlier drafts of this paper. The usual disclaimer applies.

1 Introduction

Consider a binary decision problem where one predicts whether a random outcome is a "success" ($R = 1$) or a "failure" ($R = 0$) given a continuous variable Z. Optimally, one fishes to discover the set of prediction rules that attain the highest probability of predicting success $D = 1$ when it realizes $R = 1$ (denote this conditional probability by $\beta = \Pr(D = 1|R = 1)$) given a risk that the rule predicts success when a failure realizes $R = 0$ (denote this conditional probability by $\alpha = \Pr(D = 1|R = 0)$). When Z is continuously distributed, it is straightforward (under general conditions) to generate sets of prediction rules such that each rule in a set is attached to a specific α and β. The simplest case derives from a cut-off rule that predicts success, if Z exceeds a given constant, c, and failure otherwise. By letting c vary from $-\infty$ to ∞, one obtains a set of rules that yields a specific β for any $\alpha \in [0, 1]$. The resulting plot (that depicts β as an increasing function of α) is commonly called the receiver operating characteristic (ROC) curve. It summarizes the trade-off between α and β when the prediction is based on a rule of the form $D(Z) = I(Z > c)$, where $I(E) = 1$, if E holds, and $I(E) = 0$, otherwise.

In general, the simple cut-off rule behind the ROC curve, henceforth the ROC curve rule, is not optimal. That is, another form of binary function of Z may be needed to attain the largest β given any $\alpha \in [0, 1]$. The classical Neyman-Pearson lemma implies that the optimal rule is of the form $D(Z) = I(h^*(Z) > c)$, where h^* is a monotone function of the likelihood ratio, the ratio of the distribution of Z given $R = 1$ and the distribution of Z given $R = 0$. Hence, the problem of an optimizing decision maker is to find the function h^*. We will analyze this problem under the assumption that h^* belongs to a class of continuous functions whose support may be partitioned into a finite number of intervals within which the function is either monotone increasing or monotone decreasing. Take a function h from such a class and the corresponding rule $D(Z) = I(h(Z) > c)$. By letting c vary over the range of h one can attain any $\beta \in [0, 1]$ as a monotone increasing function of $\alpha \in [0, 1]$. We call the resulting plot as the generalized ROC (GROC) curve, when h is not necessarily the optimal one, and the efficient ROC (EROC) curve, when h yields the optimal rules (i.e., $h = h^*$). As in the case of the standard ROC curve, the

generalized ROC curve summarizes the trade-off between α and β when a given prediction rule, specified by h, is applied. The EROC curve describes the maximal predictive power of the underlying continuous predictor, as it returns the largest possible β given any $\alpha \in [0, 1]$ attainable provided the efficient rules are known.

As an analog to the "area under the ROC curve" (AUROC), I define the area under the GROC curve (GAUROC) and the area under the EROC curve (EAUROC). These concepts provide summary indices for the underlying curves. More importantly, GAUROC can serve as a criterion for identifying the optimal rule among a set of rules.

We present estimation procedures for the GROC curve and the GAUROC for a given h-function and derive their asymptotic properties under the random sampling assumption. These results form the basis for empirical investigation and comparison of the performance of alternative prediction rules. We also develop procedures for estimating the efficient rules, the EROC curve and EAUROC under two settings. In the first setting, the h-function is assumed to belong to a parametric family of functions. This is regarded as a semiparametric approach, because the distribution of Z (conditional on R) remains nonparametrically specified. The estimator of the parameter of the semiparametric model is shown to be consistent (at the rate of square root of the sample size) and asymptotically normal. In the second setting, the h-function is only subject to nonparametric conditions.

The goal of the paper is to pave the way for improving binary predictions in situations where a continuous predictor is used. An important field of application is medical diagnosis, where a single "biomarker" may be used to predict whether a patient is healthy or diseased. It is standard to use the simple ROC curve rule for prediction and to apply the ROC curve to measure and compare the performance of alternative biomarkers. We point out that the ROC curve rule is not always optimal and that it may be beneficial to explore the performance of alternative rules based on the predictor. We provide well-founded concepts and tools for comparing alternative prediction rules and demonstrate that we are basically after the ROC curve rule formulated in terms of the optimal transformation of the original predictor. We provide techniques for finding the optimal transformation under semiparametric and nonparametric conditions.

While the concept of the generalized ROC curve presented in this paper is (to the

best of my knowledge) new to the literature on binary prediction and classification, its development is much inspired by McIntosh and Pepe (2002) and Pepe (2003) who point out the importance of Neyman-Pearson lemma for an optimal binary prediction. The asymptotic analysis of the empirical versions of the GROC curve and GAUROC can be regarded as an extension to the one of the empirical versions of the ROC curve and AUROC (see Hsieh and Turnbull (1996)). The proposed procedures for estimating the efficient rule are novel; I am not aware of similar estimators in the current literature. In an ongoing work, I extend some of the results of the present paper on a binary prediction problem where there are several (continuous or discrete) predictors rather than just a single continuous predictor.

Section 2 presents the prediction problem and the key concepts of the paper. Section 3 presents estimation procedures for the GROC curve and GAUROC under a large class of prediction rules. Section 4 develops estimation procedures for the efficient rule, the EROC curve and EAUROC under semiparametric and nonparametric conditions. Section 5 concludes.

2 Foundations

2.1 The Prediction Problem

The value of the binary random variable $R \in \{0, 1\}$ is predicted by using a continuous random variable Z. Call the realization $R = 1$ as "success" and $R = 0$ as "failure." The binary prediction is a decision rule $D = D(Z) \in \{0, 1\}$, an indicator function. The underlying problem is equivalent to the one of deciding whether an observation on Z is from one of two distributions, the distribution of Z conditional on $R = 0$ or the distribution of Z conditional on $R = 1$. We refer to these two distributions by the random variables X and Y. We assume that X and Y are absolutely continuous (with respect to Lebesgue measure on $\mathbb{R}$) and denote their cumulative distribution functions by F and G and density functions by f and g. By definition, X and Y are mutually independent.

Consider the conditional probabilities

$$\begin{aligned}\beta &= \Pr(D(Z)=1|R=1)=\Pr(D(Y)=1)\\ \alpha &= \Pr(D(Z)=1|R=0)=\Pr(D(X)=1)\end{aligned}$$

In biometrics, β is often called the "true positive fraction," and α the "false positive fraction" (see Pepe (2003)). In the terminology of statistical hypothesis testing, D could indicate that the "null hypothesis" ($R=0$) is "accepted" ($D=0$) or "rejected" ($D=1$) in favor of an "alternative hypothesis" ($R=1$). Hence, α can be regarded as the "size" and β as the "power" of the prediction.

Let $u(d,r)$ denote the level of "utility" (e.g., reward in terms of money) that is associated with the prediction $D=d\in\{0,1\}$ and the realization $R=r\in\{0,1\}$. Let $\Pr(R=1)=\delta\in(0,1)$ and notice that $\Pr(D=d,R=r)=\Pr(D=d|R=r)\Pr(R=r)$. One wishes to maximize the expected utility

$$\begin{aligned}E(u(D,R)) &= \beta\delta(u(1,1)-u(0,1))-\alpha(1-\delta)\left(u(0,0)-u(1,0)\right)\\ &\quad +\delta u(0,1)+(1-\delta)u(0,0)\end{aligned}\tag{1}$$

The following assumption is natural.

Assumption 1 *(i)* $u(1,1)>u(0,1)$ *and* $u(0,0)>u(1,0)$*; (ii)* $|u(d,r)|\leq M<\infty$.

By part (i) of Assumption 1 it is always better to forecast correctly. The case with $u(1,1)<u(0,1)$ and $u(0,0)<u(1,0)$ can always be returned to this case by replacing R with $\widetilde{R}=1-R$. The boundedness condition in part (ii) ensures that the expectation in (1) is well defined. If Assumption 1 does not hold, there is no interesting prediction problem. For example, if $u(1,1)>u(0,1)$ and $u(0,0)\leq u(1,0)$, it is always optimal to set $D=1$ (no matter what value Z takes).

2.2 The Efficient Prediction Rule

Under Assumption 1 it is clear that for a given α (β) one is always better off the higher β (the smaller α) is. This is as in the statistical hypothesis testing problem where one

wishes to have a test that has the highest possible power given the size of the test. The "fundamental lemma" of Neyman and Pearson (1933) gives a principle by which such an optimal test can be obtained. The optimal decision assumes the likelihood ratio (LR) rule

$$D_\alpha^*(Z) = I(LR(Z) > c) \tag{2}$$

where, for any $z \in \mathbb{R}$,

$$LR(z) = \frac{g(z)}{f(z)} \tag{3}$$

and c is a constant such that $\Pr(LR(Z) > c|R = 0) = \alpha$.

At a general level, the Neyman-Pearson lemma entails that the decision (the rule) may be randomized. Randomization is needed when $LR(z)$ is a constant over a subset of the support of Z. In this type of situation, the formula in (2) alone does not yield rules for all $\alpha \in [0, 1]$. For example, we may have $LR(z) = c$ whenever $z \in [z_1, z_2],\ z_1 < z_2$. To generate optimal rules for all possible $\alpha \in [0, 1]$ in this case, the Neyman-Pearson lemma uses a "critical function" that randomizes the decision when z takes on a value on the interval $[z_1, z_2]$. Such a critical function is not needed in the present paper, because we will below assume that $LR(z)$ is a constant at a set of measure zero, except when X and Y are equally distributed. In the exceptional case (i.e., if $F = G$), Z is independent of R and has no predictive content for R. The best that one can do in this case is to apply a purely randomized rule $D_\alpha \in \{0, 1\}$, which is independent of R and such that $\Pr(D_\alpha = d) = \alpha^d(1 - \alpha)^{1-d},\ \alpha \in [0, 1],\ d \in \{0, 1\}$. By using D_α, we get $\beta = \alpha$ for any $\alpha \in [0, 1]$.

By the Neyman-Pearson lemma, the LR rule in (2) yields the largest β for any $\alpha \in [0, 1]$. Basically, by letting the constant c in (2) vary over the range of $LR(Z)$, one obtains a continuum of rules that generate the set of points $\{(\alpha, \beta),\ \alpha, \beta \in [0, 1]\}$ such that $\beta = EF(\alpha)$ for a continuous, increasing and concave function $EF : [0, 1] \to [0, 1]$. The fact that the function EF (the "efficient frontier") is continuous, increasing and concave can be deduced by exploring the fundamental lemma (see Lehmann and Romano (2005)), but it will also become evident in Section 2.4, where we analyze the performance of rules within a general class.

In general, there may exists an efficient rule that yields $\beta = 1$ for some $\overline{\alpha} < 1$. In

this case, $EF(\alpha) = 1$ for all $\alpha \in (\overline{\alpha}, 1]$. On the other hand, if there is a rule that yields $\overline{\beta} > 0$ for $\alpha = 0$, then EF contains all of the points $\{(0, \beta),\ 0 \leq \beta \leq \overline{\beta}\}$. These types of situations are again ruled out by assumptions that we impose below. Hence, EF will be a strictly concave curve from point $(0, 0)$ to the point $(1, 1)$. Also, we typically have in mind a situation where $EF(\alpha)$ is smooth enough to be differentiable on $(0, 1)$, that is, the derivative $EF'(\alpha)$ is finite for all $\alpha \in (0, 1)$. Then, (under Assumption 1 and assuming X and Y are not equally distributed) the optimal decision rule that maximizes the expected utility in (1) is characterized by the first order condition

$$EF'(\alpha) = \frac{u(0,0) - u(1,0)}{u(1,1) - u(0,1)} \frac{1-\delta}{\delta} \tag{4}$$

Under (4), the optimum is a unique point $(\alpha, \beta) \in (0, 1)$, where the slope of $EF(\alpha)$ equals the ratio of the expected net utility of the correct prediction of failure ($R = 0$) and the correct prediction of success ($R = 1$). Letting $(u(0, 0) - u(1, 0))$, $(u(1, 1) - u(0, 1))$ or $\delta \in (0, 1)$ vary, a variety of "slopes" are possible. In particular, if $u(1, 1) - u(0, 1)$ is very small or $u(0, 0) - u(1, 0)$ is very large (δ fixed), then (4) may not hold for any $(\alpha, \beta) \in (0, 1)$, and the optimal rule boils down to a corner solution with $(\alpha, \beta) = (0, 0)$. Similarly, a corner solution with $(\alpha, \beta) = (1, 1)$ is possible, if $u(1, 1) - u(0, 1)$ is very large or $u(0, 0) - u(1, 0)$ is very small. Nevertheless, to please everybody (with any utility function) maximally, one has to be able to produce predictions that yield the efficient frontier as a whole.

Notice that given any strictly monotone increasing function $\eta : \mathbb{R} \to \mathbb{R}$, the rule in (3) is equivalent to the rule

$$D^*_\alpha(Z) = I(\eta(LR(Z)) > \eta(c))$$

In particular, as noted by McIntosh and Pepe (2002), we have the monotone relationship

$$\pi(z) = \Pr(R = 1|Z = z) = \frac{\Pr(R = 1)g(z)}{\Pr(R = 1)g(z) + \Pr(R = 0)f(z)} = \frac{\delta LR(z)}{\delta LR(z) + 1 - \delta} \tag{5}$$

and hence the rule in (2) is equivalent to the rule

$$D^*_\alpha(Z) = I(\pi(Z) > c') \tag{6}$$

where c' is such that $\Pr(\pi(Z) > c'|R = 0) = \alpha$.

Viewing the representations (2) and (6), one can observe that the optimal rules are nonparametrically (semiparametrically) identifiable either through the nonparametric (semiparametric) identification of F and G, or of $\pi(Z)$. Hence, the modeling of F and G, or of $\pi(Z)$, is a possible approach for obtaining the optimal rules. However, the modeling of F and G, or $\pi(Z)$ is often a difficult task. It is desirable to develop alternative approaches for finding the optimal rule that do not entail specifying F and G, or $\pi(Z)$. The concepts that we introduce in the following form a basis for developing such procedures, as we will demonstrate in Section 4.

2.3 A General Class of Prediction Rules

The following condition will be used in what follows.

Condition 1 *A function ℓ satisfies the condition, if its domain can be written as the union of a finite number of intervals such that over each of the intervals ℓ is strictly increasing or strictly decreasing and differentiable except possibly at the end points.*

We assume:

Assumption 2 *The density functions f and g are continuous and nonzero on $\mathbb{R}$. When R and Z are not independent, the likelihood ratio $LR(z) = g(z)/f(z)$ satisfies Condition 1.*

Under Assumption 2, when X and Y are not equally distributed, one can partition $\mathbb{R}$ into a finite number of successive intervals, within which $LR(z)$ is either strictly monotone increasing or strictly monotone decreasing. If X and Y have the same distribution (i.e., if R and Z are independent), then $LR(z) = 1$ for all $z \in \mathbb{R}$.

Given Assumption 2, our interest is to consider prediction rules of the form

$$D_c(Z) = I(h(Z) > c) \tag{7}$$

where the function h is subject to Condition 1. Under Assumption 2 the optimal rule can always be represented in the form (7) for some h that satisfies Condition 1. In the special case, where X and Y follow the same distribution, the optimal rule is purely randomized

and can be obtained by any function h meeting Condition 1. This will become evident below.

As X and Y are absolutely continuous, the transformed variables $h(X)$ and $h(Y)$ are absolutely continuous (e.g., Rohatgi 1976, p. 73). Let F_h and G_h (f_h and g_h) denote the distribution (density) functions of $h(X)$ and $h(Y)$. As f and g are nonzero on $\mathbb{R}$, the densities f_h and g_h are nonzero on A_h, the range of h, an interval.

Notice that we could well assume $A_h = \mathbb{R}$, because we can always replace h in (7) by its strictly monotone increasing transformation without changing the rule and without violating Condition 1. Similarly, without loss of generality we could modify Assumption 2 so that f and g are nonzero on a common subset of $\mathbb{R}$. If there is a set S such that $\Pr(X \in S) > 0$, while $\Pr(Y \in S) = 0$, then we know that $R = 0$, if $Z = z \in S$. There is a prediction problem only, if Z takes on a value on a set that is common to the supports of X and Y. Hence, it is not restrictive to assume that X and Y have the same support, $\mathbb{R}$.

2.4 Measuring the Performance of a Prediction Rule

Consider the set of prediction rules

$$\mathcal{D}_h = \{D_c(Z) = I(h(Z) > c),\ c \in A_h\}$$

where the subscript h signifies a particular function (meeting Condition 1) in the rule (7). For any given value $c \in A_h$ we have

$$\begin{aligned} \beta(c) &= \Pr(D_c(Y) = 1) = \Pr(I(h(Y) > c)) = 1 - G_h(c) \\ \alpha(c) &= \Pr(D_c(X) = 1) = \Pr(I(h(X) > c)) = 1 - F_h(c) \end{aligned}$$

As $F_h(c)$ and $G_h(c)$ are continuous distribution functions, $\alpha(c)$ and $\beta(c)$ are strictly monotone decreasing functions from A_h to $[0, 1]$. Consider the set

$$\{(\alpha(c), \beta(c)) = (1 - F_h(c), 1 - G_h(c)),\ c \in A_h\}$$

The inverse functions (the quantile functions) $F_h^{-1} : [0, 1] \to A_h$ and $G_h^{-1} : [0, 1] \to A_h$ are well defined. The inverse of $1 - F_h(c)$ is $F_h^{-1}(1 - t)$, a strictly monotone increasing

function from $[0,1]$ to A_h. Hence, we have

$$\{(\alpha(c), \beta(c)),\ c \in A_h\} = \{(t, GROC(t)),\ t \in [0,1]\}$$

where the function $GROC : [0,1] \to [0,1]$ is given by

$$GROC_h(t) = 1 - G_h(F_h^{-1}(1-t)) \tag{8}$$

The curve defined by the set of points $(t, GROC_h(t)), t \in [0,1]$ is called the generalized ROC (GROC) curve. It is strictly monotone increasing and expresses how β grows as a function of α, when the rule in (7) with some given function h (under Condition 1) is applied.

When h is the identity function (or any strictly monotone increasing function), the GROC curve is equal to the (standard) ROC curve:

$$ROC(t) = 1 - G(F^{-1}(1-t))$$

As an equivalent alternative to $GROC(t)$, we may consider the "generalized ordinal dominance" (GODC) curve with

$$GODC_h(t) = F_h(G_h^{-1}(t)) \tag{9}$$

which expresses $1-\alpha$ as a function of $1-\beta$, when the rule in (7) with some given function h (under Condition 1) is applied. The GODC curve reduces to the standard ODC curve ($ODC(t) = F(G^{-1}(t))$) when h is any strictly monotone increasing function.[1]

Notice that if $F = G$, then

$$GROC_h(t) = GODC_h(t) = t$$

That is, if R is independent of Z, then any h under Condition 1 yields the optimal randomized rule, with $\alpha = \beta$.

Assume $h(Z)$ is

$$h^*(Z) = \eta(LR(Z)) \tag{10}$$

[1] See Hsieh and Turnbull (1996) for discussion on the ODC curve.

where η is any strictly monotone increasing continuous function. As the rule in (7) using h from (10) is equivalent to (2), the resulting set of points $\{(\alpha(c), \beta(c)),\ c \in A_{h^*}\}$ must be optimal, and the resulting GROC curve coincides with EF (the efficient frontier). Accordingly, a GROC curve that assumes h in (10) is called the "efficient" ROC curve or the EROC curve. In this case, we write $EROC(t)$ and $\mathcal{D}^* = \{D_c^*(Z) = I(h^*(Z) > c),\ c \in A_{h^*}\}$.

Consider

$$\frac{\partial GODC(t)}{\partial t} = \frac{\partial F_h(G_h^{-1}(t))}{\partial G_h^{-1}(t)} \frac{\partial G_h^{-1}(t)}{\partial t} = \frac{f_h(G_h^{-1}(t))}{g_h(G_h^{-1}(t))} \tag{11}$$

where the last equality follows, as $\partial G_h^{-1}(t)/\partial t = 1/[\partial G_h(G_h^{-1}(t))/\partial G_h^{-1}(t)]$. From (11) one can see that the slope of the GROC curve as a function of the cut-of-point c is generally determined by the likelihood ratio $g_h(c)/f_h(c)$ of the transformed variables $h(X)$ and $h(Y)$. When $h = h^*$, this ratio is a monotone increasing function of the likelihood ratio of the original variables X and Y (i.e., $LR(c)$). As c is a monotone decreasing function of α, the slope of the $EROC$ curve is monotone decreasing, showing that the EROC curve is strictly concave (which we already known from above). In general, the GROC curve need not be concave, it can have concave and convex segments.

The following result is a consequence of the Neyman-Pearson lemma.

Theorem 1 *Assumption 2 holds. Given any $GROC_h(t)$, $EROC(t) \geq GROC_h(t)$ for all $t \in [0, 1]$.*

2.5 Summary Measures

We define the "area under the GROC curve" (GAUROC) for a given h as

$$GAUROC(h) = \int_0^1 GROC_h(t)dt$$

We have $GAUROC(h) = \int_0^1 GODC_h(t)dt$ and $GAUROC(h) \in [0, 1]$. When h is a monotone increasing function, GAUROC is equal to the area under the (conventional) ROC curve (AUROC), and we may write $AUROC = \int_0^1 ROC(t)dt$.

We define the area under the efficient ROC (EAUROC) as

$$EAUROC = \int_0^1 EROC(t)dt$$

Theorem 1 implies the following result.

Corollary 1 *Assumption 2 holds. Then* $EAUROC \geq GAUROC(h)$ *for any* h.

If $LR(z)$ is monotone increasing, then $EROC = ROC$ and $EAUROC = AUROC$. If $F = G$, then $EROC(t) = ROC(t) = GROC_h(t) = t$, and $EAUROC = AUROC = GAUROC(h) = \frac{1}{2}$ for any h under Condition 1. If $F \neq G$, then $EROC(t) > \frac{1}{2}$, while some functions h under Condition 1 yield $GAUROC(h) < \frac{1}{2}$.

We have

$$\begin{aligned} GAUROC(h) &= \int_0^1 F_h(G_h^{-1}(t))dt \\ &= \int_{A_h} F_h(s)dG_h(s) \\ &= \int_{A_h} \left(\int_{-\infty}^{s} f_h(u)du \right) g_h(s)ds \\ &= \int_{A_h} \int_{A_h} I(u < s) f_h(u) g_h(s) du ds \\ &= \Pr(h(X) < h(Y)) \end{aligned}$$

where we apply the change of variable formula (to replace t by $s = G_h^{-1}(t)$) and the last line follows from the independence of X and Y. Notice that when h is a monotone increasing function, $GAUROC = \Pr(X < Y) = AUROC$, where the last equality is well known from the ROC curve literature. As is also well known, the so called Mann and Whitney U-statistic for testing whether $F = G$ is an estimate of $\Pr(X < Y)$.

An alternative summary index for a GROC curve is defined as

$$KS(h) = \max_t |GROC_h(t) - t|$$

It can be shown that $KS(h)$ is equivalent to the Kolmogorov-Smirnov measure of distance between the distributions F_h and G_h. When h is a monotone increasing function, i.e., if $GROC_h(t) = ROC(t)$, then $KS(h) = KS = \max_t |ROC(t) - t|$. Pepe (2003) calls KS (we call $KS(h)$) as the Kolmogorov-Smirnov ROC (GROC) measure.

Define for a given rule h its "maximal hit rate" as

$$MHR(h) = \max_c \Pr(D(h(Z) > c) = R)$$

The following result is obvious.

Corollary 2 *Assumption 2 holds. Then* $KS(h^*) \geq KS(h)$ *and* $MHR(h^*) \geq MHR(h)$ *for any* h.

Each of the summary indices, $GAUROC(h)$, $KS(h)$, or $MHR(h)$ is maximized by the EROC curve and can be used as a criterion for identifying the efficient rule (for more on identification see Section 4.1). In particular, note that the efficient rule is identified as the one yielding the best "hit rate," i.e., the minimum percentage of wrong predictions. Other measures of distance between F_h and G_h (such as the Cramer von Mises criterion) could be used as a basis for summarizing the GROC curve, and such measures would also identify the efficient rule.

2.6 An Illustration

Assume X (Z given $R = 0$) follows the standard normal distribution, and Y (Z given $R = 1$) follows the extreme value distribution with the location parameter and the scale parameter both equal to 1. The density functions of X and Y are shown in Figure 1.

The prediction rule (7) based on a given function h is conveniently expressed by the representation

$$D_\alpha(Z; h) = I(1 - F_h(h(Z)) < \alpha) \tag{12}$$

where F_h is the distribution function of $h(X)$ and $\alpha = \Pr(D_\alpha(X; h) = 1) \in [0, 1]$. The optimal decision rule is based on $h = h^*$, where $h^*(z)$ is any strictly monotone increasing function of the likelihood ratio $LR(z)$. In the present example, we can write h^* as

$$h^*(z) = z + \frac{z^2}{2} - \exp(z - 1) \tag{13}$$

The function $h^*(z)$ is depicted in Figure 2(a), while Figure 2(b) shows the conditional probability function, $\Pr(R = 1|Z = z)$. Figure 2(c) shows the function $a^*(z) = 1 - F_{h^*}(h^*(z))$ and illustrates the optimal rule by means of the presentation in (12). When $\alpha = 0.1$, the optimal rule is $D^*_{0.1}(Z) = I(Z < -2.62) + I(1.29 < Z < 2.75)$. This is shown by the blue lines in Figure 2(c). Similarly, the red lines in Figure 2(c) indicate how $D^*_{0.5}(Z)$ is determined. The blue and red lines in Figures 2(a) and 2(b) correspond to the ones of Figure 2(c).

Figure 2(d) displays the EROC curve (the red curve) that is obtained by using the formula in (8) with $h = h^*$. The blue solid line is the conventional ROC curve based on the rule with h being the identity function $h(z) = z$ (and $D_\alpha(Z; h) = I(1 - F(Z) < \alpha)$), while the blue dotted line (denoted as "ROC(−1)") is obtained by applying the ROC curve rule on the $-Z$, i.e., $h(z) = -z$. The efficient rule is superior to the two ROC curve rules, but coincides with the "negative" ROC curve rule, when α is close to zero ($\alpha < 0.0018$) or one ($\alpha > 0.9992$). The value of EAUROC is 0.671, while the AUROC based on Z is 0.64 and the one based on $-Z$ is 0.36 ($= 1 - 0.64$).

3 Estimation Under a Given Rule

In applied work one wishes to estimate the GROC curve for alternative rules. For example, one may want to assess empirically whether a "quadratic rule," $D(Z^2 > c)$, has better performance than the conventional ROC curve rule, $D(Z > c)$. This section presents basic estimation theory for the GROC curve of any given rule. In addition, we introduce the empirical counterpart of GAUROC and analyze its asymptotic properties.

3.1 Empirical GROC Curve

The GROC curve of any given prediction rule $D(h(Z) > c)$ can be estimated in the same manner as the standard ROC curve. There are two types of sampling settings. In a "two-sample setting," we have two independent samples, one on X, another on Y. In a "one-sample setting," we have a single random sample on (R, Z). In both cases, we denote observations on X (or Z given $R = 0$) by $X_1, ..., X_{n_X}$ and those on Y (or Z given $R = 1$) by $Y_1, ..., Y_{n_Y}$. In this section, following the convention of the ROC curve literature (see Hsieh and Turnbull (1996)), we assume the two-sample setting and present asymptotic results under $n = n_X + n_Y \to \infty$ such that $n_Y/n \to \kappa \in (0, 1)$. With some minor modification, the asymptotic results hold also under the one-sample setting. In the one sample setting, we have $n \to \infty$, and due to random sampling $n_Y/n \to \delta = \Pr(R = 1)$.

We seek to estimate $GROC_h(t) = 1 - G_h(F_h^{-1}(1 - t))$ or equivalently $GODC_h(t) = F_h(G_h^{-1}(t))$ for a given function h. The standard empirical counterparts for F_h and G_h

are given by

$$\widehat{F}_h(u) = \frac{1}{n_X}\sum_{i=1}^{n_X} I(h(X_i) \le u) \text{ and } \widehat{G}_h(u) = \frac{1}{n_Y}\sum_{i=1}^{n_Y} I(h(Y_i) \le u)$$

while the inverses F_h^{-1} and G_h^{-1} are estimated by $\widehat{F}_h^{-1}(s) = \inf(u : \widehat{F}_h(w) \ge s)$ and $\widehat{G}_h^{-1}(s) = \inf(u : \widehat{G}_h(w) \ge s)$. We estimate $GROC_h$ and $GODC_h$, respectively, by

$$\widehat{GROC}_h(t) = 1 - \widehat{G}_h(\widehat{F}_h^{-1}(1-t)) \text{ and } \widehat{GODC}_h(t) = \widehat{F}_h(\widehat{G}_h^{-1}(t))$$

It suffices to consider the latter estimator.

Theorem 2 *Assumption 2 holds. h meets Condition 1. (a)*

$$\sup_{t\in[0,1]} \left|\widehat{GODC}_h(t) - GODC_h(t)\right| \xrightarrow{a.s.} 0, \text{ as } n \to \infty$$

(b) There exists a probability space on which one can define sequences of two independent Brownian bridges $\{B_n^1(t), 0 \le t \le 1\}$, $\{B_n^2(t), 0 \le t \le 1\}$ such that on any subinterval $[a,b]$ of $(0,1)$ on which $f_h(G_h^{-1}(t))/g_h(G_h^{-1}(t))$ is bounded, we have

$$\begin{aligned}\sqrt{n}\left(\widehat{GODC}_h(t) - GODC_h(t)\right) &= \sqrt{\frac{1}{1-\kappa}}B_n^1(F_h(G_h^{-1}(t))) \\ &\quad + \sqrt{\frac{1}{\kappa}}\frac{f_h(G_h^{-1}(t))}{g_h(G_h^{-1}(t))}B_n^2(t) + o_p(n^{-\frac{1}{2}}\log(n)^2) \text{ a.s.}\end{aligned}$$

uniformly on $[a,b]$, as $n \to \infty$.

If h is the identity function (i.e., when $GODC = ODC$ and $GROC = ROC$), Theorem 2 agrees with Theorems 2.1 and 2.2 of Hsieh and Turnbull (1996). Theorem 2 follows from arguments given in Hsieh and Turnbull (1996).

3.2 Empirical GAUROC

For a given h, the obvious estimator of $GAUROC(h) = \Pr(h(X) < h(Y))$ is

$$\widehat{GAUROC}(h) = \int_0^1 \widehat{F}_h(\widehat{G}_h^{-1}(t))dt = \frac{1}{n_X n_Y}\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} I(h(X_i) < h(Y_j)) \qquad (14)$$

Analogously to Hsieh and Turnbull (1996), we can apply Theorem 2 to show that $\widehat{GAUROC}(h)$ is consistent and asymptotically normal. Here we present more general limiting results that concern a set of h-functions jointly.

Assume the two-sample setting as in Section 3.1. Notice that $\widehat{GAUROC}(h)$ for any given h is a two-sample U-statistics with kernel $I(h(X_i) < h(Y_j))$ specified by h. Let $\mathcal{H}$ denote a class of h-functions. Then, the family $\left\{\widehat{GAUROC}(h), h \in \mathcal{H}\right\}$ is a two-sample U-process and the corresponding standardized two-sample U-process is $\{U_{n_X n_Y}(h), h \in \mathcal{H}\}$, where

$$U_{n_X n_Y}(h) = \sqrt{n}\left(\widehat{GAUROC}(h) - \Pr(h(X) < h(Y)\right)$$

An application of the central limit theorem for two-sample U-processes of Neumeyer (2004) yields the following result.

Theorem 3 *Assumption 2 holds. $\mathcal{H}$ is a class of functions that meet Condition 1. Then, as $n \to \infty$, the process $\{U_{n_X n_Y}(h), h \in \mathcal{H}\}$ converges weakly to a Gaussian process $\{\mathcal{G}(h), h \in \mathcal{H}\}$ with zero mean and covariance kernel*

$$\begin{aligned}\operatorname{cov}(\mathcal{G}(h_1), \mathcal{G}(h_2)) &= \frac{1}{1-\kappa}\int_{-\infty}^{\infty}\left[1 - G_{h_1}(h_1(x))\right]\left[1 - G_{h_2}(h_2(x))\right] f(x)dx \\ &\quad + \frac{1}{\kappa}\int_{-\infty}^{\infty} F_{h_1}(h_1(y))F_{h_2}(h_2(y))g(y)dy \\ &\quad - \frac{1}{(1-\kappa)\,\kappa}\left(\int_0^1 F_{h_1}(G_{h_1}^{-1}(t))dt\right)\left(\int_0^1 F_{h_2}(G_{h_2}^{-1}(t))dt\right)\end{aligned}$$

Theorem 3 implies:

Corollary 3 *Assumption 2 holds. Take a given h under Condition 1. Then,*

$$\sqrt{n}\left(\widehat{GAUROC}(h) - \Pr(h(X) < h(Y))\right) \xrightarrow{d} N(0, V_h), \text{ as } n \to \infty$$

where

$$\begin{aligned}V_h &= \frac{1}{1-\kappa}\int_0^1 \left[1 - G_h(F_h^{-1}(t))\right]^2 dt + \frac{1}{\kappa}\int_0^1 F_h(G_h^{-1}(t))^2 dt \\ &\quad - \frac{1}{\kappa(1-\kappa)}\left(\int_0^1 F_h(G_h^{-1}(t))dt\right)^2\end{aligned}$$

When h is a monotone increasing function, we have $\widehat{GAUROC}(h) = \widehat{AUROC}$ and Corollary 3 agrees with Theorem 2.3 of Hsieh and Turnbull (1996).

Note that in Corollary 3 we can write

$$V_h = \frac{1}{1-\kappa}\int_0^1 GROC_h(t)^2 dt + \frac{1}{\kappa}\int_0^1 GODC_h(t)^2 dt - \frac{1}{\kappa(1-\kappa)} GAUROC(h)^2$$

The following uniform convergence in probability follows from Theorem 2.9 of Neumeyer (2004).

Theorem 4 *The conditions of Theorem 3 hold. Then*

$$\sup_{h\in\mathcal{H}} \left| \widehat{GAUROC}(h) - \Pr(h(X) < h(Y)) \right| = o_p(1)$$

Under the one-sample setting, the result in Theorem 4 can be strengthened to uniform almost sure convergence (see Neumeyer (2004, p. 78)).

4 Finding and Estimating the Efficient Rule

The primary goal of an optimizing decision maker is to find the efficient rules, which allow her to choose the optimum point as in Section 2.2. The previous section shows that one can consistently estimate the GROC curve for any given rule. This allows searching for the best rule among any alternative rules, but in practice one wishes to have a procedure that automatically finds the efficient rule. We consider this problem in semi-parametric and non-parametric settings.

4.1 The Starting Point

The starting point of the estimation of the efficient rules is that one specifies a class of rules

$$\mathcal{C} = \{\mathcal{D}_h,\ h \in \mathcal{H}\}$$

where each $\mathcal{D}_h$ is as in (7) and $\mathcal{H}$ is a given family of functions that meet Condition 1. We configure the set $\mathcal{H}$ so that its members amount to distinct, unique rules. The following condition is sufficient.

Condition 2 *The family of functions $\mathcal{H}$ is such that the set of roots $\{z : h(z) = c,\ c \in A_h\}$ are real and unique to each $h \in \mathcal{H}$.*

Under Condition 2, $\mathcal{H}$ can be interpreted as an index set for a class of rules. Let h^* denote the efficient rule and assume $h^* \in \mathcal{H}$. We estimate h^* by

$$\widehat{h} = \arg\max_{h \in \mathcal{H}} \widehat{GAUROC}(h) \tag{15}$$

In Section 4.2, we analyze (15) in a situation where $\mathcal{H}$ is a parametric class of functions. This is regarded as a semiparametric estimation approach, because it does not entail specifying F and G or $\Pr(R|Z)$ parametrically. Section 4.3 advances a procedure for handling (15) when $\mathcal{H}$ is specified nonparametrically.

The general motivation of the estimator in (15) arises from Corollary 1[2], that is,

$$GAUROC(h^*) \geq GAUROC(h) \text{ for all } h \in \mathcal{H} \tag{16}$$

By Condition (16) and Theorem 4, one can show that the estimator $\widehat{h}$ is consistent for h^*.[3] Here we must recognize that h^* is not always unique (even if Condition 2 holds). That is, there may be a set $H^* \subset \mathcal{H}$ of several (or a "continuum" of) rules such that $GAUROC(h) = EAUROC$ for all $h \in H^*$ and $EAUROC > GAUROC(h)$ for all $h \in H$, where $H^* \cup H = \mathcal{H}$ and $H^* \cap H = \varnothing$. Any h in H^* is efficient and yields the EROC curve. In this situation, the estimator $\widehat{h}$ is consistent in the sense that it picks up one of such rules with probability tending to one, as $n \to \infty$. Hence, $\widehat{h}$ is a reasonable estimator even if h^* is not unique. Nevertheless, situations where h^* is not unique require some attention when the estimator is applied in practice. Two mains cases are discussed in the following.

The first case arises when $F = G$, as then any h (under Condition 1) is efficient and we have $H^* = \mathcal{H}$ and $H = \varnothing$. The problem in (15) is not very interesting, as any solution

[2]Given Corollary 1, the estimation problem in (15) could be based on alternative criterion functions such as $KS(h)$ or $MHR(h)$, but such alternatives are not studied in this paper.

[3]Heuristically, as n tends to infinity, we eventually learn $GAUROC(h)$ for all $h \in \mathcal{H}$ and thereby we find h^* that maximizes $GAUROC(h)$, $h \in \mathcal{H}$. Formally, we must restrict $\mathcal{H}$ to be a compact metric space and such that h^* is an inner point of the space. One relevant metric is $d(h, h') = |GAUROC(h) - GAUROC(h')|$. We would then have $d(\widehat{h}, h^*) = o_p(1)$. We leave a careful probability theoretical treatment of the estimator for later research.

to it is efficient. Hence, for the estimation problem in (15) to be meaningful, we must have $F \neq G$ (i.e., R cannot be independent of Z). Sometimes it is not clear in advance whether R depends on Z. One approach for handling this question is to apply existing tests for the hypothesis $F = G$.

The second case arises when h^* is monotone. While the class of monotone functions under Condition 2 is uncountable, we only need to examine whether h^* is increasing or decreasing. The problem in (15) is very simple. We choose $\widehat{h}(z) = z$, if $\widehat{AUROC} \geq 0.5$, and $\widehat{h}(z) = -z$, otherwise. The problem in (15) becomes more interesting when we suspect that h^* is a nonmonotone function. Hence, in applied work, it would be useful to have a procedure for investigating whether h^* is monotone or not. If one finds evidence that h^* is not monotone, then (under Assumption 1) one can ask whether there are one or more "turning points," where h^* switches between increasing and decreasing segments. The procedures that we will consider below assume that one knows the maximum number of turning points of h^*. When $\mathcal{H}$ is a parametric class of functions (Section 4.2), then this information is incorporated into the chosen parametric class of functions. In the case of the nonparametric procedure (Section 4.3), this is the key restriction imposed on the class $\mathcal{H}$.

Suppose h^* is a non-monotone function and that we have specified $\mathcal{H}$ so that $h^* \in \mathcal{H}$. Then under our assumptions there is generally a single function h^* for which

$$GAUROC(h^*) > GAUROC(h) \text{ for all } h \neq h^*,\ h \in \mathcal{H} \tag{17}$$

That is, $GAUROC(h)$ identifies a unique optimal rule. This is desirable at least from the point of view of conventional estimation theory. However, there are situations where the identification condition in (17) does not quite bite in the sense that there may be a set of rules that are "locally" efficient even if they are "globally" inefficient. As we demonstrate in the following section, such a situation can arise, when F and G are both normal. In this example, h^* is in general a quadratic function. However, it is possible that the probability mass of Z concentrates almost solely on the increasing (decreasing) segment of h^* so that there is only a negligible probability that Z takes on a value at the decreasing (increasing) segment of h^*. It is then possible that one cannot identify the true optimal rule even with

a large number of observations. This type of situation may be a nuisance for statistical inference, but the estimator $\widehat{h}$ is still working to the right direction and captures in large samples a rule that is effectively equal to the efficient rule.

Finally, in applied work it is of course possible that the assumed class $\mathcal{H}$ is misspecified in the sense that $h^* \notin \mathcal{H}$. In such a situation, maximizing $GAUROC(h)$ over $h \in \mathcal{H}$ does not yield the best rule for all situations (i.e., for all α). This is because the GROC curves of two different rules can cross. Nevertheless, it is still possible that there is a single rule $h^\dagger \in \mathcal{H}$ such that

$$GAUROC(h^\dagger) > GAUROC(h) \text{ for all } h \neq h^\dagger,\ h \in \mathcal{H}$$

Hence, the maximization of $GAUROC(h)$ may well identify some rule even if $h^* \notin \mathcal{H}$. While the identified rule is not efficient in this case, it may be a good approximation to the efficient one, as we will demonstrate below.

4.2 Semi-parametric Approach

One possible choice for $\mathcal{H}$ in (15) is a parametric family of functions. We can think that the underlying parametric function is fully specified by a parameter vector ψ of dimension $p+1$. To meet Condition 2, ψ must be subject to a constraint or a normalization. Without loss of generality, we can assume that once such constraint is imposed on ψ we can write $\psi = (\theta_0, \theta')'$, where $\theta_0 \in \mathbb{R}$ is fixed and $\theta \in \Theta \subseteq \mathbb{R}^p$. As a result, individual functions within the parametric family are indexed by θ. Accordingly, write h as h_θ, $h(z)$ as $h(z;\theta)$. Furthermore, we write F_θ (G_θ) for F_h (G_h), $GROC_\theta(t)$ for $GROC_h(t)$, $GAUROC(\theta)$ for $GAUROC(h)$, and so on. Suppose $h_\theta = h^*$ for some $\theta = \theta^*$. The problem in (15) reduces to

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \widehat{GAUROC}(\theta) \tag{18}$$

Section 4.2.1 gives an illustration when $\mathcal{H}$ constitutes a set of polynomial functions. Section 4.2.2 shows that $\widehat{\theta}$ is $\sqrt{n}$-consistent and asymptotically normal under regularity conditions.

4.2.1 Polynomial Rules

A rule is called a "polynomial rule," if it can be expressed by

$$h(z;\psi) = \psi_1 z + \psi_2 z^2 + \cdots + \psi_{p+1} z^{p+1} \tag{19}$$

In order to meet Condition 2, this function must be "scale normalized." We may impose $\|\psi\| = 1$ or $\psi_j = 1$ for some $j \in \{1, 2, ..., p+1\}$. If we choose $\psi_1 = 1$ (as we do below), then our parameter of interest is $\theta = (\theta_1, \theta_2, ..., \theta_p)' = (\psi_2, \psi_3, ..., \psi_{p+1})'$, while $\theta_0 = \psi_1 = 1$ is fixed.

When we set $p = 0$ in (19), we obtain a "linear rule." This rule is solely specified by the sign of ψ_1 (the value of ψ_1 does not matter). If $\psi_1 > 0$ (e.g., $\psi_1 = 1$), we have a ROC curve rule, and if $\psi_1 < 0$ (e.g., $\psi_1 = -1$), we have a "negative ROC curve rule" (as it was named earlier). If $LR(z)$ ($h^*(z)$) is monotone, then one of these rules is efficient and we identify it as the one that maximizes AUROC. If $LR(z)$ is non-monotonic, then one of the rules maximizes AUROC, but none of the two is efficient. If $F = G$, (i.e., if R and Z are independent), then both of the rules are efficient, as they can serve as a purely randomized rule. Finally, in this setting, the estimation problem in (19) boils down to one, where we choose $\widehat{h}(z) = z$, if $\widehat{AUROC} \geq 0.5$, and $\widehat{h}(z) = -z$, otherwise.

Setting $p = 1$ in (19) yields the "quadratic rule"

$$h(z;\psi) = \psi_1 z + \psi_2 z^2 \tag{20}$$

If X is standard normal and Y is normal with mean μ and variance σ^2, it is easy to see that the efficient rule is obtained by setting $\psi_1 = \mu/\sigma^2$ and $\psi_2 = (\sigma^2 - 1)/2\sigma^2$. If X and Y have unequal means ($\mu \neq 0$), but the same variance ($\sigma^2 = 1$), then the ROC curve rule (with $\psi_1 = \text{sign}(\mu)$) is efficient. Otherwise, the efficient rule assumes a quadratic polynomial (we must have $\psi_2 \neq 0$). Notice that when $\mu = 0$ and $\sigma^2 = 1$, i.e., when $F = G$, the above formulae yield $\psi_1 = \psi_2 = 0$. Nonetheless, in this situation, any nonzero values of ψ_1 and ψ_2 yield a purely randomized rule, which is efficient when R is independent of Z. Finally, note that the quadratic rule does not entail that X and Y (or $T(X)$ and $T(Y)$ for some common transformation function T) are normal.

Normalize $\psi_1 = 1$ in (20) so that the rule can be written as

$$h(z;\theta) = z + \theta z^2 \tag{21}$$

where $\theta = \psi_2/\psi_1$. Keep with the above example, i.e., $X \sim N(0,1)$ and $Y \sim N(\mu, \sigma^2)$. Then the efficient rule is given by $\theta = \theta^* = (\sigma^2 - 1)/2\mu$. Let $\mu = 1$. Figure 3 displays GAUROC as a function of θ when (a) $\sigma^2 = 4$, $\theta^* = 1.5$, (b) $\sigma^2 = 1.5$, $\theta^* = 0.25$, (c) $\sigma^2 = 0.6$, $\theta^* = -0.2$, (d) $\sigma^2 = 0.1$, $\theta^* = -0.45$. In each case, $GAUROC(\theta)$ is a smooth concave function and attains its maximum at θ^*. In cases (a) and (d), $GAUROC(\theta)$ serves as a clear (population level) criterion for identifying θ^*. However, when θ^* is closer to 0 (or when σ is closer to 1), $GAUROC(\theta)$ is fairly flat around θ^*. In particular, in case (c), one can hardly recognize that $GAUROC(\theta)$ is maximized at $\theta^* = -0.2$. These observations indicate that when θ^* is nonzero, but close to 0, it is difficult to distinguish the efficient rule from the simple ROC curve rule (obtained by setting $\theta = 0$). It also turns out that in these cases the ROC curve rule does not lose much (if anything) compared to the efficient rule. An illustration follows.

Figure 4 displays the actual densities f and g as well as the function $1 - F_{\theta^*}(z) = \alpha$ (describing the efficient rule) and the function $1 - F(z) = \alpha$ (describing the ROC curve rule) for the cases of Figure 3. The dotted black horizontal lines (the green solid lines) describe the efficient rule (the ROC curve rule) for $\alpha = 0.1$. For example, in panel (b) of Figure 4, the efficient rule for $\alpha = 0.1$ is

$$I(1 - F_{\theta^*}(Z) < 0.1) = I(Z < -5.2816 \text{ or } Z > 1.2816)$$

while the ROC curve rule for $\alpha = 0.1$ is

$$I(1 - F(Z) < 0.1) = I(Z > 1.2816)$$

Observe from the underlying densities that the "lower triggering condition" $Z < -5.2816$ of the efficient rule "kicks in" extremely rarely (for example, if $\Pr(R = 1) = 0.5$, we have $\Pr(Z < -5.2816) = 1.049 \times 10^{-7}$). On the other hand, the "upper triggering condition" $Z > 1.2816$ of the efficient rule is the same as the one of the ROC curve rule. That is, the two rules do not differ in practice when $\alpha = 0.1$. Overall, one can see that in the case

(b) the efficient rule and the ROC curve rule are practically identical whenever α is not larger than, say, 0.8. For larger values of α (i.e., if $\alpha \in (0.8, 1)$), one can see from the figure that the "lower triggering condition" of the efficient rule hits with some recognizable probability and that the "lower triggering condition" of the efficient rule differs slightly from the one of the ROC curve rule. Hence, when α is close to 1, the efficient rule and the ROC curve rule differ to the extent that it might matter in some practical application.

Along similar lines one can see from Figure 4 that in the case (c) the efficient and the ROC curve rule are virtually identical. For example, when $\alpha = 0.1$, the efficient rule is

$$I(1 - F_{\theta^*}(Z) < 0.1) = I(1.281 < Z < 3.719)$$

while the ROC curve rule is

$$I(1 - F(Z) < 0.1) = I(Z > 1.282)$$

The two rules do not differ much in practice, because it is rare that $Z > 3.719$ (for example, if $\Pr(R = 1) = 0.5$, we have $\Pr(Z > 3.719) = 0.0079$). The same conclusion holds over the whole range of $\alpha \in (0, 1)$. By contrast when one looks at cases (a) and (d) in Figure 4 one can recognize that the efficient rule and the ROC curve rule are clearly different for most values of α.

Figure 5 depicts the EROC curve and the ROC curve for the cases of the previous figures. As one can expect, in cases (a) and (d), the EROC curve is superior to the ROC curve for a large range of values of α, while in cases (b) and (c) one can hardly recognize that the EROC curve is above the ROC curve. Figure 6 shows the actual difference $ROC(t) - EROC(t)$ between the curves in each case. The magnitude of the difference is recognizable in cases (a) and (d), but negligible in cases (c) and (d).

The above example illustrates that in some situations a "continuum" of rules are practically indistinguishable from the efficient rule. In such situations, $GAUROC(\theta)$ is nearly flat around θ^*, which makes it difficult to identify and estimate the "exact" efficient rule in any finite sample. However, as the example illustrated, in such situations, any rule sufficiently close to the efficient rule is optimal in practical terms. Hence, there are situations where it suffices that one is able to pick up a rule from a large set of equally

good rules. This is what $h(\widehat{\theta})$ tends to do, because it is (as a special case of $\widehat{h}$ in (15)) consistent for h^*.

In the appendix we analyze the performance of the polynomial rule in the example of Section 2.6, where X is standard normal and Y follows the extreme value distribution. In this situation the true efficient h^* cannot be written as a polynomial (of any finite order). Hence, the assumed family of rules is misspecified in the sense that it does not include the efficient rule. However, the polynomial rule is shown to yield a very good approximation to the efficient rule. The example illustrates that polynomial rules may have merit in a variety of situations.

4.2.2 Asymptotic Distribution

Assume the one-sample setting. Denote the underlying random sample on (R, Z) by $(R_1, Z_1), ..., (R_n, Z_n)$. Notice that the estimation problem in (18) remains the same when $\widehat{GAUROC}(\theta)$ is replaced by

$$C_n(\theta) = \frac{1}{n(n-1)} \sum_{i \neq j} I(R_i > R_j) I(h(Z_i; \theta) > h(Z_j; \theta)) \tag{22}$$

This criterion function is similar in form to the one of the maximum rank correlation (MRC) estimator of Han (1987) with the exception that in the MRC estimator $h(Z; \theta)$ is replaced by a linear combination of exogenous regressors. If $h(Z; \theta)$ is a polynomial of order p, then $C_n(\theta)$ corresponds to the criterion function of the MRC estimator, where the regressors are the polynomial terms Z^j, $j = 1, ..., p$. Sherman (1993) presents general methods that can be applied to show that the MRC estimator is $\sqrt{n}$-consistent and asymptotically normal. The methods handle the fact that the criterion function in (22) is not differentiable with respect to θ. To apply these methods in the present setting, we impose the following assumption.

Assumption 3 *Let $\mathcal{N}$ denote a neighborhood of θ^*.*
(i) For each z in $\mathbb{R}$, all mixed second order partial derivatives of $F_\theta(h(z; \theta))$ and of $G_\theta(h(z; \theta))$ exist on $\mathcal{N}$.
(ii) There is an integrable function $M(z)$ such that for all z in $\mathbb{R}$ and θ in $\mathcal{N}$ it holds that

$\left\| \frac{\partial^2}{\partial\theta\partial\theta'} F_\theta(h(z;\theta)) \right\| \leq M(z)|\theta - \theta^*|$ *and* $\left\| \frac{\partial^2}{\partial\theta\partial\theta'} G_\theta(h(z;\theta^*)) \right\| \leq M(z)|\theta - \theta^*|$.
(iii) $E\left|\frac{\partial}{\partial\theta} F_\theta(h(Z;\theta^*))\right|^2 < \infty$ *and* $E\left|\frac{\partial}{\partial\theta} G_\theta(h(Z;\theta^*))\right|^2 < \infty$.
(iv) $E\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j} F_\theta(h(Z;\theta^*))\right| < \infty$ *and* $E\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j} G_\theta(h(Z;\theta^*))\right| < \infty$, $i,j = 1,...,p$.
(v) The matrix $E\left\{\frac{\partial^2}{\partial\theta\partial\theta'}\left[(1-\delta)I(R=1)F_\theta(h(Z;\theta^*)) + \delta I(R=0)\left(1 - G_\theta(h(Z;\theta^*))\right)\right]\right\}$ *is negative definite.*

Theorem 5 *Let the efficient parameter* θ^* *be an interior point of* Θ*, a compact subspace of* $\mathbb{R}^p$*. Then, under Assumption 3, as* $n \to \infty$, $\sqrt{n}(\widehat{\theta} - \theta^*) \xrightarrow{d} N(0, J^{-1}QJ^{-1})$, *where*

$$\begin{aligned}(\delta - \delta^2)^{-1}Q &= (1-\delta)E\left\{\left[\frac{\partial}{\partial\theta}F_\theta(h(Y;\theta^*))\right]\left[\frac{\partial}{\partial\theta}F_\theta(h(Y;\theta^*))\right]'\right\} \\ &\quad +\delta E\left\{\left[\frac{\partial}{\partial\theta}G_\theta(h(X;\theta^*))\right]\left[\frac{\partial}{\partial\theta}G_\theta(h(X;\theta^*))\right]'\right\}\end{aligned}$$

and

$$V = (\delta - \delta^2)\frac{\partial^2}{\partial\theta\partial\theta'}GAUROC(\theta^*)$$

Note that condition (iii) of Assumption 3 implies that the matrix Q exists. The fact that the matrix V exists and is non-singular follows from condition (v) of Assumption 3. In fact, in condition (v), it is legitimate to change the order of expectation (integration) and differentiation so that one can see that the matrix in the condition is equal to V (see the proof of the theorem). Basically, the role of Assumption 3 is to ensure that the expectation of the criterion function $C_n(\theta)$ can be approximated sufficiently accurately by a second order Taylor expansion around θ^*. This is one of the key conditions applied to handle the fact that $C_n(\theta)$ is not differentiable with respect to θ.

Given the estimate $\widehat{\theta}$, we can estimate $EROC_\theta(t)$ by $\widehat{GROC}_{\widehat{\theta}}(t)$ and $EAUROC_\theta$ by $\widehat{GAUROC}(\widehat{\theta})$. As $\widehat{\theta}$ is consistent for θ^* we obtain the following result.

Corollary 4 *The conditions of Theorem 5 hold. Then, as* $n \to \infty$, $\widehat{GAUROC}(\widehat{\theta}) \xrightarrow{p} EAUROC$ *and* $\sup_{t\in[0,1]}\left|\widehat{GROC}_{\widehat{\theta}}(t) - EROC(t)\right| \xrightarrow{p} 0$.

The quantities Q and V in Theorem 5 can be expressed alternatively as in the binary choice model example of Sherman (1993). First, note that $h(Z;\theta^*) = \eta(LR(Z))$ for some strictly increasing function η. Let η^{-1} be the inverse function of η and note that by (5)

we have

$$\Pr(R = 1|Z) = \Lambda(h(Z;\theta^*))$$

where $\Lambda(\cdot) = \delta\eta^{-1}(\cdot)/(\delta\eta^{-1}(\cdot)+1-\delta)$ is a strictly increasing distribution function. Hence, the binary variable R is determined by

$$R = I(h(Z;\theta^*) - \varepsilon > 0) \tag{23}$$

where ε is a random variable (error) that is independent of Z and has the distribution function Λ. This representation is similar to the binary choice model analyzed in detail by Sherman (1993) with the exception that in Sherman (1993) $h(Z;\theta)$ is replaced by a weighted sum of some regressors, where the weights are determined by θ.

Write U^* for $\partial h(Z;\theta^*)/\partial\theta$ and $\overline{U}^*$ for $E(U^*|h(Z;\theta^*))$. Let $q^*_\theta(w)$ ($q_\theta(w)$) denote the density of $h(Z;\theta^*)$ ($h(Z;\theta)$) and let $\lambda(t) = \partial/\partial t\Lambda(t)$. Now, provided the assumptions of Theorem 5 hold, we can write

$$Q = E\left\{(U^* - \overline{U}^*)(U^* - \overline{U}^*)' \left[q^*_\theta(h(Z;\theta^*))\right]^2 \Lambda(h(Z;\theta^*))[1 - \Lambda(h(Z;\theta^*))]\right\}$$

and

$$V = -E\left\{(U^* - \overline{U}^*)(U^* - \overline{U}^*)' q^*_\theta(h(Z;\theta^*))\lambda(h(Z;\theta^*))\right\}$$

These expressions correspond to "Δ" and "V" in page 134 of Sherman (1993).

Suppose $\widehat{\theta}$ is an estimator for the quadratic rule $h(z;\theta) = z + \theta z^2$, when X is $N(0,1)$ and Y is $N(\mu,\sigma^2)$, as in the example of Section 4.2.1. The quantities Q and V (based on either of the above presented expressions) can be derived explicitly for any θ. It is found that whenever $\mu \neq 0$ and $\theta \neq 0$ ($\sigma^2 \neq 1$), Q and V are well defined. That is, for the estimator $\widehat{\theta}$ (multiplied by $\sqrt{n}$) to have a finite asymptotic variance, the underlying two normal distributions must have different means and variances. Recall that under this condition both coefficients in (20) are non-zero. Figure 7 plots the asymptotic standard deviation ($\sqrt{Q/V^2}$) of $\widehat{\theta}$ for a range of values of $\theta^* = (\sigma^2 - 1)/2\mu$, when $\mu = 1$. The asymptotic standard deviation of $\widehat{\theta}$ is extremely large whenever θ^* is at most about 0.1 away from 0 (alternatively when σ^2 is at most about 0.2 away from 1). This observation does not mean that $\widehat{\theta}$ yields an unreliable estimate for the efficient rule. It reflects the fact

that when θ^* is close to zero, $GAUROC(\theta)$ is very flat around θ^* so that a large range of values of θ yield GROC curves that are indistinguishable from the EROC curve. For example, when $\theta^* = 0.1$ ($\sigma^2 = 1.2$), the difference $EAUROC - GAUROC(\theta)$ is less than 0.0000001 for all $\theta \in [-.11, 0)$ and all $\theta \in (0, 0.15]$. Such a small difference is possible only when the corresponding GROC curve is extremely close to the EROC curve. Simulations indicate that in this type of situations, even in quite large samples $\widehat{\theta}$ tends to concentrate on a value from such a region rather than on θ^*. The sample size must be extremely large for $\widehat{\theta}$ to concentrate on the exact θ^*, as the consistency of $\widehat{\theta}$ entails, as $n \to \infty$.

The above discussed example reveals that there are situations where θ^* is so poorly identified, even if it is a unique maximizer of $GAUROC(\theta)$, that $\widehat{\theta}$ tends to behave as if it were consistent for a value from a set of values around θ^*. Fortunately, the underlying set is such that values within it yield rules that hardly ever deviate from the efficient rule. However, these types of cases have implications for statistical inference. For example, it would be of interest to test whether $\theta^* = 0$, i.e., whether the ROC curve rule or a monotone rule is efficient. In the above example, the estimator $\widehat{\theta}$ does not offer a reliable basis for such a test. The reason is that a range of values of θ is able to approximate the efficient rule very accurately "locally." Here "local" refers to the part of the support of Z where Z takes on values most of the time. In the above example, if $\theta^* = 0$ ($\sigma^2 = 1$), then about 99.9% of the probability mass of Z lays within the interval $[-3.1, 4.1]$. Now, the quadratic rule $h(z;\theta) = z + \theta z^2$ is monotone increasing within this range whenever $\theta \in (0, 0.16]$. Hence, a quadratic rule $h(z;\theta) = z + 0.16z^2$ yields efficient predictions on average 999 out of 1000 times. One can imagine that a huge number of observations is required for one to be able to recognize a difference between any $\theta \in (0, 0.16]$ and the efficient parameter $\theta^* = 0$.

While the fact that the criterion function in (18) (equivalently (22)) is not continuous is no concern for the estimation theory, it causes some trouble for the practical implementation. A trick by which one can avoid discrete optimization is to apply a modified criterion function where the discontinuous function $I(h(Z_i;\theta) > h(Z_j;\theta))$ is replaced by its "smoothed" version $K((h(Z_i;\theta) > h(Z_j;\theta))/\varsigma_n)$, where K is a symmetric distribution function with a continuous second derivative and $\varsigma_n > 0$ is a decreasing number satisfying

$\varsigma_n \to 0$, as $n \to 0$ (see Ma and Huang 2007). The continuous function $K(z/\varsigma_n)$ is better approximation to $I(z)$, the larger n is. The smoothing of the criterion function does not alter the asymptotic properties of the estimator.

4.3 Non-parametric Approach

In this section we consider the estimator in (15) when $\mathcal{H}$ is a nonparametric class of functions that satisfy Assumption 2. That is, our starting point is that $LR(Z)$ is a continuous function consisting of a finite number, henceforth denoted by s_{LR}, of decreasing and increasing segments. Assumption 2 entails that there is a finite s_{LR}. Here we make the additional assumption that we know a maximum $s_{\max}$ (a finite integer) such that $s_{LR} \leq s_{\max}$.

Recall from Section 2.6 that a rule based on h can be expressed as

$$D_\alpha(Z;h) = I(a(Z) < \alpha) \tag{24}$$

where

$$a(z) = 1 - F_h(h(z))$$

Any h-function is related to a unique a-function (as a is a strictly monotone transformation of h). If h meets Condition 1 so does a. Corresponding to a class of h-functions there is always an equivalent class of a-functions. The problem of finding the efficient $h(z)$, $h^*(z)$, is the same as the one of finding the efficient $a(z)$, $a^*(z) = 1 - F_{h^*}(h^*(z))$.

Let $\mathcal{A}_s$ denote the class of a-functions that satisfy Condition 1 with the qualification that the number of switches between decreasing and increasing segments is exactly s. For a given a-function let $\widehat{GAUROC}(a)$ be as $\widehat{GAUROC}(h)$ in (14) with h replaced by $-a$ (also, an h-function). Then, given a known $s_{\max}$, we estimate the efficient a^* (and hence the efficient rule) by

$$\widetilde{a} = \arg \max_{a \in \{\widetilde{a}_0, \widetilde{a}_1, ..., \widetilde{a}_{s_{\max}}\}} \widehat{GAUROC}(a) \tag{25}$$

where

$$\widetilde{a}_s = \arg \max_{a \in \mathcal{A}_s} \widehat{GAUROC}(a), \;\; s = 0, 1, ..., s_{\max} \tag{26}$$

If $s_{LR} \leq s_{\max}$, then $\widetilde{a}$ is consistent for a^* (as was discussed in Section 4.1). In what follows, we illustrate how the problem (25) can be solved in practice.

Consider the problem of finding $\widetilde{a}_s$, i.e., solving (26) for a given s. If $s = 0$, then a is either (i) increasing, or (ii) decreasing. If $s = 1$, then a is either (i) first increasing, then decreasing, or (ii) first decreasing, then increasing. If $s = 2$, a is either (i) first increasing, then decreasing, then increasing, or (ii) first decreasing, then increasing, then decreasing. That is, for a given s, a is either 'first increasing' (property (i)), or 'first decreasing' (property (ii)). If a has property (i), then $-a$ (the "mirror image" of a) has property (ii). Let $\overline{\mathcal{A}}_s = \{a : a \in \mathcal{A}_s \text{ and } a \text{ has property (i)}\}$ and $\underline{\mathcal{A}}_s = \{a : -a \in \overline{\mathcal{A}}_s\}$. We have $\mathcal{A}_s = \overline{\mathcal{A}}_s \cup \underline{\mathcal{A}}_s$ and $\overline{\mathcal{A}}_s \cap \underline{\mathcal{A}}_s = \varnothing$. We can solve (26) as follows. Find $\overline{a}_s$ and $\underline{a}_s$, respectively, that maximizes and minimizes $\widehat{GAUROC}(a)$ over $a \in \overline{\mathcal{A}}_s$. Then, we have $\widetilde{a}_s = \overline{a}_s$, if $\widehat{GAUROC}(\overline{a}_s) > 1 - \widehat{GAUROC}(\underline{a}_s)$, and $\widetilde{a}_s = \underline{a}_s$ otherwise.[4]

Assume the case $s_{\max} = 1$. We need to find $\widetilde{a}_0$ and $\widetilde{a}_1$, of which $\widetilde{a}_0$ is straightforward (you only need to check whether $\widehat{AUROC} > 0.5$). Consider $\widetilde{a}_1$. The rules corresponding to $a \in \overline{\mathcal{A}}_1$ are of the form

$$D_\alpha(Z) = I(Z < \tau_1(\alpha) \text{ or } Z > \tau_2(\alpha)) \tag{27}$$

where the node τ_1 (τ_2) is increasing (decreasing) function of α with $\tau_1(1) = \tau_2(1)$, $\tau_1(0) = -\infty$, $\tau_2(0) = \infty$. Clearly, if $a \in \overline{\mathcal{A}}_1$, then the corresponding "node function" $\tau_1(\alpha)$ ($\tau_2(\alpha)$) is the inverse function of $a(z)$ when $a(z)$ is increasing (decreasing). The problem of finding $\overline{a}_1$ can be stated as the one of finding $\tau_1(\alpha)$ and $\tau_2(\alpha)$ such that the underlying empirical GAUROC is maximized. Let $Z_{(j)}, j = 1, ..., n$ denote a sample of observations on Z and assume $Z_{(j)}$ are all unequal and ranked such that $Z_{(j)} < Z_{(j+1)}$ for all j. The rule in (27) results in the same classification of the observations whenever $\tau_1 \in (Z_{(\underline{j}-1)}, Z_{(\underline{j})}]$ and $\tau_2 \in [Z_{(\overline{j})}, Z_{(\overline{j}+1)})$ for $\underline{j}, \overline{j}$ such that $1 < \underline{j} \leq \overline{j} < n - 1$. Clearly, there is a finite number of possible classifications and these are obtained by all possible choices of $\tau_1 = Z_{(\underline{j})}$ and $\tau_2 = Z_{(\overline{j})}$ ($1 < \underline{j} \leq \overline{j} < n - 1$). A classification is specified by a pair $(\underline{j}, \overline{j})$ of observation ranks and results in corresponding estimates $(\widehat{\alpha}(\underline{j}, \overline{j}), \widehat{\beta}(\underline{j}, \overline{j}))$ of (α, β). There are more

[4] In practice, one may be able to conclude that $\widetilde{a}_s = \overline{a}_s$ ($\widetilde{a}_s = \underline{a}_s$) without solving for $\underline{a}_s$ ($\overline{a}_s$). This is seen in the simulated example discussed at the end of this section (see footnote 5).

(rank index) pairs $(\underline{j}, \overline{j})$ than different estimates $(\widehat{\alpha}, \widehat{\beta})$. An empirical rule is specified by a sequence of pairs $J = \{(\underline{j}_k, \overline{j}_k),\ \underline{j}_k, \overline{j}_k \in \{1, ..., n\},\ \underline{j}_k \leq \underline{j}_{k+1},\ \overline{j}_k \geq \overline{j}_{k+1},\ \underline{j}_k \leq \overline{j}_k\}$. A sequence J and the associated "threshold observations" $Z_{(\underline{j})}$ and $Z_{(\overline{j})}$ $((\underline{j}, \overline{j}) \in J)$ amount to corresponding empirical node functions $\tau_1(\widehat{\alpha})$, $\tau_2(\widehat{\alpha})$ and estimates for the GROC curve and GAUROC (henceforth denoted as $\widehat{GAUROC}(J)$). The estimate $\overline{a}_1$ is obtained by maximizing $\widehat{GAUROC}(J)$ over all possible J.

To further illustrate the problem of finding $\overline{a}_1$, suppose X is standard normal and Y is normal with mean $\mu = 1$ and variance σ^2. As was seen in Section 4.2, if $\sigma^2 \neq 1$, we obtain the efficient rule by $h^*(Z) = Z + \theta^* Z^2$, where $\theta^* = (\sigma^2 - 1)/2$. Alternatively, we can write

$$h^*(Z) = (Z - \varrho)^2$$

where $\varrho = -1/(\sigma^2 - 1)$. Henceforth, assume $\sigma^2 = 4$ so that $\varrho = -1/3$. As h^* has the property (ii), the corresponding a^* has property (i), and the optimal rule is of the form (27). The optimal function $a^*(z) = 1 - F_{h^*}(h^*(z))$ is blotted as a black line in Figure 8. Note that $\tau_1^*(\alpha)$ $(\tau_2^*(\alpha))$ is the inverse of $a^*(z)$ when $z \in (-\infty, -1/3]$ $(z \in [-1/3, \infty))$. The blue solid line (the red solid line) depicts the density f (g) of X (Y).

The blue squares (the red crosses) in Figure 8 are associated with observations on X (Y) based on a simulated sample (assuming the above specified setting together with $\Pr(R = 1) = 0.6$) of size $n = 100$ $(n_X = 45,\ n_Y = 55)$. The ticks of the vertical axis indicate the alpha estimates that can be obtained by using the rule (27) on the simulated observations. Squares and crosses along the same vertical line refer to a single observation, but for each tick we mark only observations that can yield the corresponding estimate $\widehat{\alpha}$. For example, the estimate $\widehat{\alpha} = 0$ can be attained by choosing $\tau_1 \leq X_{(1)}$ and $\tau_2 \geq X_{(n)}$, and we have $\widehat{\alpha} > 0$, if $\tau_1 > X_{(1)}$ or $\tau_2 < X_{(n)}$. The black circled observations and the associated dashed line constitute the estimate, $\widetilde{a}$, of the efficient rule a^*. That is, the rule yields the largest GAUROC estimate among rules consistent with (27) and the ROC curve rule. Visually the estimated rule is fairly similar to the efficient rule. The GAUROC estimate $\widehat{GAUROC}(\widetilde{a}) = 0.780$ is not much larger than $EAUROC = 0.744$. The green circled observations and the associated dashed line show the empirical ROC curve rule in the present sample.

Figure 9 shows all estimates $(\widehat{\alpha}, \widehat{\beta})$ (the black dots) obtainable from the simulated sample by using either the rule in (27) or the ROC curve rule. The red solid line in the figure is the EROC curve. The black line with circles is the estimated EROC curve, i.e., the empirical GROC curve based on the estimate of the efficient rule (the black dotted line with circles in Figure 8).[5] The estimated EROC curve gives a good approximation to the true EROC curve. Qualitatively, it is also equally accurate as the empirical GROC curve based on the true efficient rule that is shown as the red line with circles. This suggests that the nonparametric estimate of the efficient rule is about as accurate as one can hope given the available sample. Simulations confirm that nonparametric estimates for the efficient rule, EROC and EAUROC become more accurate the larger is the sample size.

The above described estimation procedure can be easily extended to cases with $s_{\max} > 1$. For example, when $s_{\max} = 2$, the set of possible rules can be expressed by

$$D_\alpha(Z) = I(Z < \tau_1(\alpha) \text{ or } \tau_2(\alpha) < Z < \tau_3(\alpha)) \tag{28}$$

where $\tau_1(\alpha) \leq \tau_2(\alpha) \leq \tau_3(\alpha)$ for all $\alpha \in [0,1]$. If $a \in \overline{\mathcal{A}}_2$, then τ_1 (τ_2) [τ_3] is increasing (decreasing) [increasing] function of α, and there are constants $\alpha^{(1)}$, $\alpha^{(2)}$, $\alpha^{(3)}$ and $\alpha^{(4)}$ such that $\alpha^{(1)} < \alpha^{(2)}$, $\alpha^{(2)} > \alpha^{(3)}$, $\alpha^{(3)} < \alpha^{(4)}$, $\tau_1(\alpha^{(2)}) = \tau_2(\alpha^{(2)})$, $\tau_2(\alpha^{(3)}) = \tau_3(\alpha^{(3)})$, $\tau_1(\alpha^{(1)}) = -\infty$, $\tau_3(\alpha^{(4)}) = \infty$, and $\alpha^{(j)} = 0$ for $j = 1$ or $j = 3$ or both, and $\alpha^{(j)} = 1$ for $j = 2$ or $j = 4$ or both. If $a \in \overline{\mathcal{A}}_1$, then we set $\tau_3(\alpha) = \infty$ for all $\alpha \in [0,1]$ so that (28) reduces to (27). Finally, the case of the ROC curve rule, $a \in \overline{\mathcal{A}}_0$, is obtained from (28) by setting $\tau_2(\alpha) = \tau_3(\alpha)$ for all $\alpha \in [0,1]$. We have already seen how $\overline{a}_0$ and $\overline{a}_1$ are obtained. The case $\overline{a}_2$ makes no essential difference to $\overline{a}_1$. A rule is specified by a sequence of threshold triples $\tau_1 = Z_{(j_1)}$, $\tau_2 = Z_{(j_2)}$ and $\tau_3 = Z_{(j_3)}$ $(1 < j_1 \leq j_2 \leq j_3 < n-1)$. A finite number of such sequences can be formed within the sample and the problem is to find the one that maximizes the corresponding empirical GAUROC. Similar procedures apply to any finite $s_{\max}$.

In this section, we have shown how (15) can be solved when the underlying class

[5] It is easy to conclude from the figure that $\widehat{GAUROC}(\overline{a}_1) > 1 - \widehat{GAUROC}(\underline{a}_1)$ so that there is no need to solve for $\underline{a}_1$.

$\mathcal{H}$ is as general as Condition 1 allows. We use the above derivations to show that the different sets generated by a rule under Condition 1 forms a "VC class," after Vapnik and Chervonenkis (1971), or a "polynomial class," after Pollard (1984). That is, we show that the number of classifications of the observations induced by a rule under Condition 1 grows at most at a rate n^k (for some finite k) that is much smaller than the maximum number of classifications, 2^n. Clearly, when $a \in \overline{\mathcal{A}}_0$, the number of possible classifications is given by the number of possible thresholds, the n observations. When $a \in \overline{\mathcal{A}}_1$, a given threshold $\tau_1 = Z_{(j)}, j = 1, ..., n-1$ can be combined with at most $n-1$ thresholds $\tau_2 \in \{Z_{(j+1)}, ..., Z_{(n)}\}$ so that there are at most n^2 classifications. It is easy to see that when $a \in \overline{\mathcal{A}}_s$, there are at most n^s classifications. It follows that the sets generated by a rule based on $a \in \overline{\mathcal{A}}_s$ is a VC class for any finite s. A complement of a VC class is a VC class and a finite union of VC classes is a VC class (Pollard (1984)). Hence, the sets generated by a rule based on $a \in \mathcal{A} = \{\mathcal{A}_0 \cup \mathcal{A}_1 \cup \cdots \cup \mathcal{A}_{s_{\max}}\}$ forms a VC class. An implication is that various function classes that appear in the proofs of Theorems 3, 4 and 5 can be shown to be "Euclidian," as argued in the proofs.

5 Conclusion

The ROC curve is a standard device for measuring the predictive power of a single continuous variable for a binary outcome. It is pointed out that the ROC curve assumes a specific prediction rule that is optimal only if the underlying likelihood ratio is monotone increasing, or in other words, if larger values of the predictor are always associated with larger probability of success. Such an assumption may be reasonable in certain applications, but cannot hold in general. For example, a potential predictor (like blood pressure) of the health of a patient can indicate lower risk of disease over a middle range of values, and higher risk when the predictor takes either on a very small or a very large value. For such and more complicated settings, a variety of alternative prediction rules can potentially improve upon the simple ROC curve rule. The generalized ROC curve of the paper describes the performance of a given rule and allows one to make comparisons between alternative rules.

It is also pointed out that there is always an efficient prediction rule that allows one to exploit the maximal predictive power of a predictor. Anyone's interest is to find the efficient rule as it beats all competing rules no matter how one's utilities with respect to different outcomes are configured. Importantly, the paper offers novel procedures that allow one to estimate the efficient rule under general semiparametric and nonparametric conditions. The nonparametric conditions entail that the underlying likelihood ratio is smooth enough and such that it alters between monotone decreasing and increasing segments. It is plausible that this condition is general enough to capture the efficient rule in applications.

The proposed estimation procedures call for further development. The nonparametric estimator for the efficient rule assumes that one knows the maximum number of switches between increasing and decreasing segments of the likelihood ratio. As such information may be uncertain in practice, it is desirable to develop procedures for estimating the exact number of turning points of the likelihood ratio. In particular, it would be useful to have a test for whether the underlying likelihood ratio is monotone and thereby whether the popular ROC curve rule is efficient. An example in the paper shows that this problem can be cumbersome due to potential poor identification of the efficient rule under the null. It is of interest to extend the present work to situations were there are several predictors. This is one of extensions that the author is currently working on.

Appendix

Proof of Theorem 3

The result follows from Theorem 2.7 of Neumeyer (2004). Notice that $U_{n_X n_Y}(h)$ is analogous to "$U_{nm}(f)$" of Neumeyer (2004). Define the conditional expectations

$$\begin{aligned} E_X(h)(y) &= E(I(h(X) < h(Y))|Y = y) \\ E_Y(h)(x) &= E(I(h(X) < h(Y))|X = x) \end{aligned}$$

and write these as $E_X(h)$ and $E_Y(h)$, when the value of the conditioning variable is not fixed. Let

$$E(h) = E(I(h(X) < h(Y)))$$

Here $E_X(h)$, $E_Y(h)$, $E(h)$ correspond to the notations Pf, Qf, $P \otimes Q(f)$ of Neumeyer (2004). Then, given Theorem 2.7 of Neymeyer, the covariance stated in Theorem 3 is equal to

$$\frac{1}{1-\kappa} E_X\left[(E_Y(h_1))(E_Y(h_2))\right] + \frac{1}{\kappa} E_Y\left[(E_X(h_1))(E_X(h_2))\right] - \frac{1}{(1-\kappa)\kappa} E(h_1)E(h_2)$$

We have

$$\begin{aligned} E_Y(h)(x) &= E(I(h(x) < h(X))) = 1 - G_h(h(x)) \\ E_X(h)(y) &= E(I(h(X) < h(y))) = F_h(h(y)) \end{aligned}$$

and hence it is immediate that $E_X\left[(E_Y(h_1))(E_Y(h_2))\right]/(1-\kappa)$ is the same as the first term in covariance of Theorem 3. Similar arguments show that the two remaining terms above agree with the ones of the covariance of Theorem 3.

For the conditions of Theorem 2.7 of Neumeyer (2004), it suffices that the class of functions $\{I(h(X) < h(Y)), h \in \mathcal{H}\}$ and the classes defined by the conditional expectations $E_X(h)$ and $E_Y(h)$ (for $h \in \mathcal{H}$) are all "Euclidian," see Neumeyer (2004, p. 79). For the required property it suffices that the sets generated by $I(h(X) < h(Y))$ form a VC class of sets (Vapnik and Chervonenkis (1971)), or the polynomial class of sets in the terminology of Pollard (1984). The VC property follows from Condition 1 on the h-function. For a detailed argument, see the end of Section 4.3.

Illustrating Semiparametric Model under Misspecification

Recall the example of Section 2.6, where X is standard normal and Y follows the extreme value distribution. The associated efficient h-function is given in (13) and we observe that no monotone function of it can be represented as a polynomial. To investigate how well the quadratic rule "approximates" the efficient one in this case, consider

$$h_2(z;\theta) = z + \theta z^2$$

The maximum value of GAUROC is 0.6694 and is obtained with $\theta = 0.64$. The GAUROC maximizing value of θ is fairly close to the coefficient (0.5) of z^2 in the efficient function in (13). Figure A1 shows that the quadratic rule results in a good approximation to the optimal rule in the sense that its GROC curve is very close to the EROC curve. One can hardly recognize that the GROC curve is below the EROC curve except when α is close to 0. This is rather surprising given that the efficient rule assumes two "turning points," while the quadratic rule has only one turning point. A rule based on a third order polynomial $h_3(z;\theta) = z + \theta_1 z^2 + \theta_2 z^3$, a "cubic rule," can produce two turning points. The corresponding GROC curve, with parameter values that maximize GAUROC among all cubic rules, is shown as a blue line in Figure A1. One can see that the cubic rule is better than the quadratic rule when α is small. However, as Figure A2 shows, the cubic rule is not uniformly better than the quadratic rule.

Keeping with the example of Section 2.6, Figure A3 illustrates how the (GAUROC maximizing) linear (ROC), quadratic and cubic rule track the efficient rule. It is helpful to note that the different rules can be expressed as

$$D_\alpha(Z;h) = I(Z < \tau_1(\alpha;h)) + I(\tau_2(\alpha;h) < Z < \tau_3(\alpha;h))$$

where the nodes $\tau_j(\alpha;h)$ are specific to h and satisfy $\tau_1(\alpha;h) \le \tau_2(\alpha;h) \le \tau_3(\alpha;h)$. Let $h = h_1$, $h = h_2$, $h = h_3$, respectively, for the linear, the quadratic and the cubic rule and write $\tau_j(\alpha;h^*) = \tau_j^*(\alpha)$ (for the efficient rule). When $\alpha \in (0.00181, 0.9993)$, the "efficient nodes" $\tau_j^*(\alpha)$, $j = 1,2,3$ are all distinct and finite ($\tau_j^*(0.6)$ are marked in Figure A3). When $\alpha \in (0, 0.00181]$, $\tau_2^*(\alpha) = \tau_3^*(\alpha)$, while when $\alpha \in [0.9993, 1)$, $\tau_1^*(\alpha) = \tau_2^*(\alpha) = -\infty$. The performance of a polynomial rule depends on how well its nodes $\tau_j(\alpha;h_i)$ track the "efficient nodes" $\tau_j^*(\alpha)$. The linear (the negative ROC) and the cubic rules agree with the efficient rule when $\alpha \in [0, 0.00181)$ and also when $\alpha \in (0.9993, 1]$ (the linear rule) or when $\alpha \in (0.9998, 1]$ (the cubic rule). The quadratic rule assumes $\tau_3(\alpha;h_2) = \infty$ for all α. Despite the fact that $\tau_3(\alpha;h_2)$ is "infinitely far" from $\tau_3^*(\alpha)$ (for all finite $\alpha < 1$), the quadratic rule outperforms the cubic rule over a large set $\alpha \in [0.5658, 0.9656]$ (see Figure A2). This can be explained by observing from Figure A3 that for the underlying values of α, $\tau_1(\alpha;h_2)$ and $\tau_2(\alpha;h_2)$ follow $\tau_1^*(\alpha)$ and $\tau_2^*(\alpha)$ much closer than $\tau_1(\alpha;h_3)$ and $\tau_2(\alpha;h_3)$

do. In particular, $\tau_1(\alpha; h_2)$ and $\tau_2(\alpha; h_2)$ are close to the efficient nodes within a range where Z has a large mass (see Figure 1). It does not matter so much that $\tau_3(\alpha; h_2)$ is infinitely far apart from $\tau_3^*(\alpha)$, if $\tau_3^*(\alpha)$ is in a region where Z has very little probability mass (as then it is rare that $Z > \tau_3^*(\alpha)$).

Proof of Theorem 5

Write $W = (R, Z)$. W takes values on the set $S = \{0, 1\} \otimes \mathbb{R}$. Denote the underlying random sample by $W_i = (R_i, Z_i)$, $i = 1, ..., n$.

For each (w_1, w_2) in $S \otimes S$, and each θ in Θ, define the function

$$\psi(w_1, w_2, \theta) = I(r_1 > r_2) I(h(z_1; \theta) > h(z_2; \theta))$$

Write the criterion function as

$$C_n(\theta) = \frac{1}{n(n-1)} \sum_{i \neq j} \psi(W_i, W_j, \theta)$$

As was noted earlier, $C_n(\theta)$ can be seen as a U-statistic of order two, and the collection $\{C_n(\theta), \theta \in \Theta\}$ is a U-process of order two. This is as in the framework of Sherman (1993).

We can write the conditional expectation of $\psi(W_1, W_2, \theta)$ given $W_1 = w = (r, z)$ as

$$\begin{aligned} E\left[\psi(w, W, \theta)\right] &= E\left[(r > R) I(h(z; \theta) > h(Z; \theta))\right] \\ &= \Pr(R = 0) I(r > 0) E\left[I(h(X; \theta) < h(z; \theta))\right] \\ &= (1 - \delta) I(r = 1) F_\theta(h(z; \theta)) \end{aligned}$$

Similarly,

$$E\left[\psi(W, w, \theta)\right] = \delta I(r = 0)(1 - G_\theta(h(z; \theta))$$

Define the function

$$\tau(w, \theta) = (1 - \delta) I(r = 1) F_\theta(h(z; \theta)) + \delta I(r = 0)(1 - G_\theta(h(z; \theta))$$

We have

$$\begin{aligned} E(C_n(\theta)) &= E\left[\psi(W_1, W_2, \theta)\right] \\ &= \delta(1 - \delta) E\left[h(Y; \theta) > h(X; \theta)\right] \\ &= \delta(1 - \delta) GAUROC(\theta) = C(\theta) \end{aligned}$$

and

$$\begin{aligned} E(\tau(W;\theta)) &= (1-\delta)\delta E\left[F_\theta(h(Y;\theta))\right] + \delta(1-\delta)E\left[(1-G_\theta(h(X;\theta))\right] \\ &= 2\delta(1-\delta)GAUROC(\theta) = 2C(\theta) \end{aligned}$$

Now, applying the "U-statistic decomposition" of Sherman (1993, p. 127), we can write

$$C_n(\theta) = C(\theta) + \frac{1}{n}\sum_i \upsilon(W_i,\theta) + \frac{1}{n(n-1)}\sum_{i\neq j} e(W_i,W_j,\theta)$$

where

$$\upsilon(w,\theta) = \tau(w,\theta) - 2C(\theta)$$

and

$$e(w_1,w_2,\theta) = \psi(w_1,w_2,\theta) - E\left[\psi(w_1,W,\theta)\right] - E\left[\psi(W,w_2,\theta)\right] + C(\theta)$$

Below, we will show that conditions equivalent to those in Assumptions A1-A4 of Sherman (1993, p. 129) hold. Theorem 5 then follows from Theorem 4 of Sherman (1993, p. 129). The quantity "Δ" of Sherman (1993) is here given by

$$\begin{aligned} & E\left\{\left[\frac{d}{d\theta}\tau(W;\theta^*)\right]\left[\frac{d}{d\theta}\tau(W;\theta^*)\right]'\right\} \\ = \; & (1-\delta)^2\delta E\left\{\left[\frac{\partial}{\partial\theta}F_\theta(h(Y;\theta^*))\right]\left[\frac{\partial}{\partial\theta}F_\theta(h(Y;\theta^*))\right]'\right\} \\ & +(1-\delta)\delta^2 E\left\{\left[\frac{\partial}{\partial\theta}G_\theta(h(X;\theta^*))\right]\left[\frac{\partial}{\partial\theta}G_\theta(h(X;\theta^*))\right]'\right\} = Q \end{aligned}$$

and the quantity "V" of Sherman (1993) is here

$$\begin{aligned} & E\left(\frac{\partial^2}{\partial\theta\partial\theta'}\tau(W,\theta^*)\right)/2 \\ = \; & E\left[(1-\delta)I(R=1)\frac{\partial^2}{\partial\theta\partial\theta'}F_\theta(h(Z;\theta^*)) + \delta I(R=0)\frac{\partial^2}{\partial\theta\partial\theta'}(1-G_\theta(h(Z;\theta^*)))\right]/2 \\ = \; & \frac{\partial^2}{\partial\theta\partial\theta'}\left\{(1-\delta)E\left[I(R=1)F_\theta(h(Z;\theta^*))\right] + \delta E\left[I(R=0)(1-G_\theta(h(Z;\theta^*)))\right]\right\}/2 \\ = \; & \frac{\partial^2}{\partial\theta\partial\theta'}\left\{(1-\delta)\delta E\left[F_\theta(h(Y;\theta^*))\right] + (1-\delta)\delta E\left[1-G_\theta(h(X;\theta^*))\right]\right\}/2 \\ = \; & (1-\delta)\delta\frac{\partial^2}{\partial\theta\partial\theta'}GAUROC(\theta^*) = V \end{aligned}$$

We have assumed that θ^* is an interior point of Θ, a compact set. This is as in A1 of Sherman (1993). A2 of Sherman (1993) is implied by the presentation in (23). The

role of A3 of Sherman (1993) is to ensure that the term "$X'\beta_0$" in his paper is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}$. Here, "$X'\beta_0$" is replaced by $h(Z;\theta^*)$, which is absolutely continuous by conditions we have imposed. Sherman (1993) notes that the consistency of his MRC estimator was established by Han (1987) and that Han's proof used the compactness of Θ, A2 and A3. In the present setting, Theorem 4 implies that $\widehat{\theta}$ is consistent for θ^*, as θ^* is an interior point of Θ and as θ^* is a unique maximizer of $GAUROC(\theta)$ (and hence of $C(\theta)$) over Θ. Finally, the conditions of Assumption 3 are equivalent to those of A4 of Sherman (1993).

References

Han, A. K. (1987) "Non-parametric analysis of a generalized regression model," *Journal of Econometrics*, 35, 303–316.

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

Lehmann, E. L., and J. P., Romano (2005) "Testing statistical hypotheses (3rd ed.)," Springer, New York, USA.

Ma, S., and J., Huang (2007) "Combining multiple markers for classification using ROC," *Biometrics*, 63, 751–757.

McIntosh, M. W., and M. S., Pepe (2002) "Combining several screening tests: optimality of the risk score," *Biometrics*, 58, 657–664.

Neyman, J., and E. S. Pearson (1933) "On the problem of the most efficient tests of statistical hypothesis," *Philosophical Transactions of the Royal Society of London, Series A*, 231, 289–337.

Neumeyer, N. (2004) "A central limit theorem for two-sample U-processes," *Statistics and Probability Letters*, 67, 73–85.

Pepe, J. L. (2003) "The statistical evaluation of medical tests for classification and prediction," Oxford University Press, Oxford, UK.

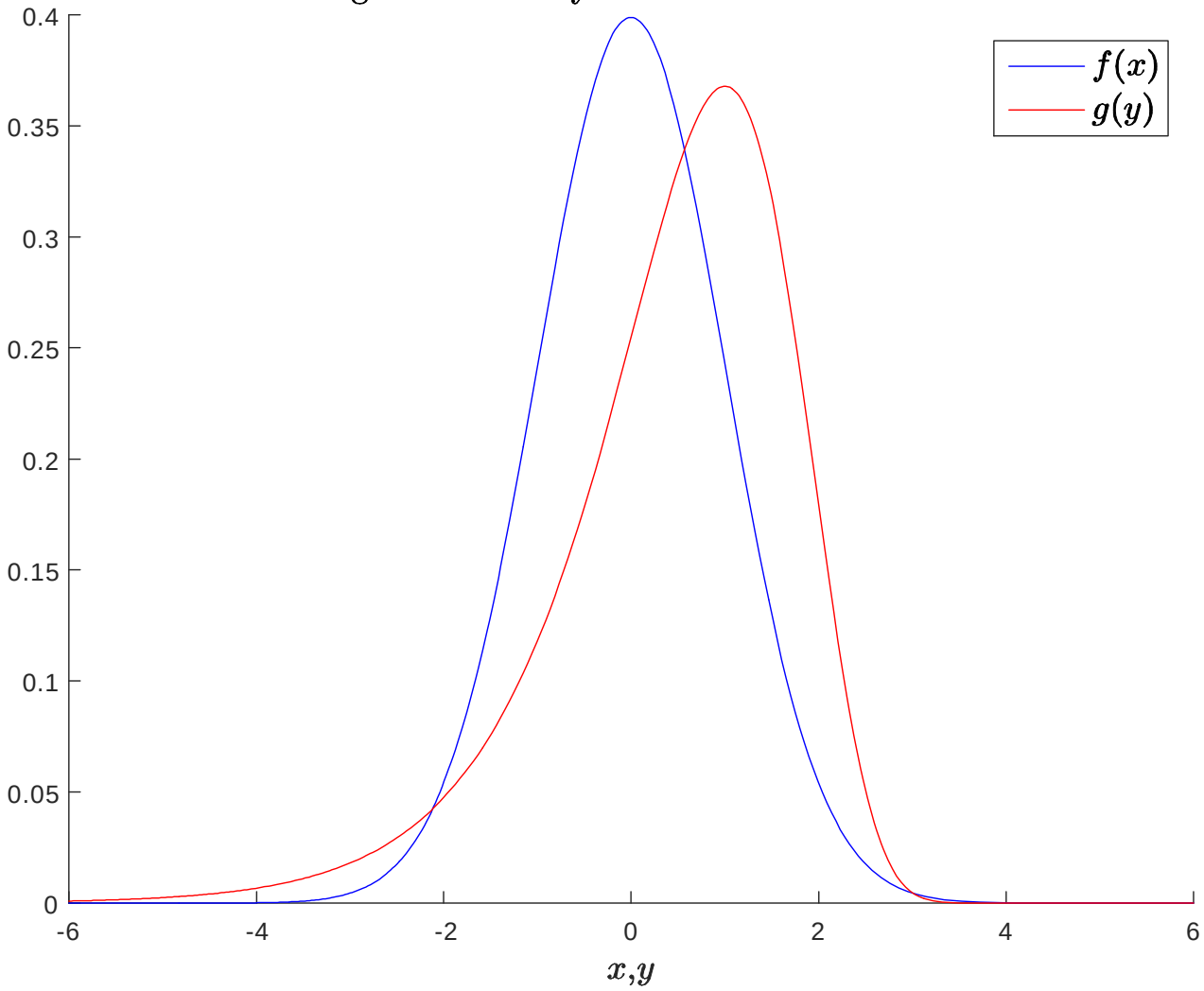
Pollard, D. (1984) "Convergence of stochastic processes," Springer, New York, USA.

Rohatgi, V. K. (1976) "An introduction to probability theory and mathematical statistics," John Wiley & Sons, New York, USA.

Sherman, V. K. (1993) "The limiting distribution of the maximum rank correlation estimator," *Econometrica*, 61, 123–137.

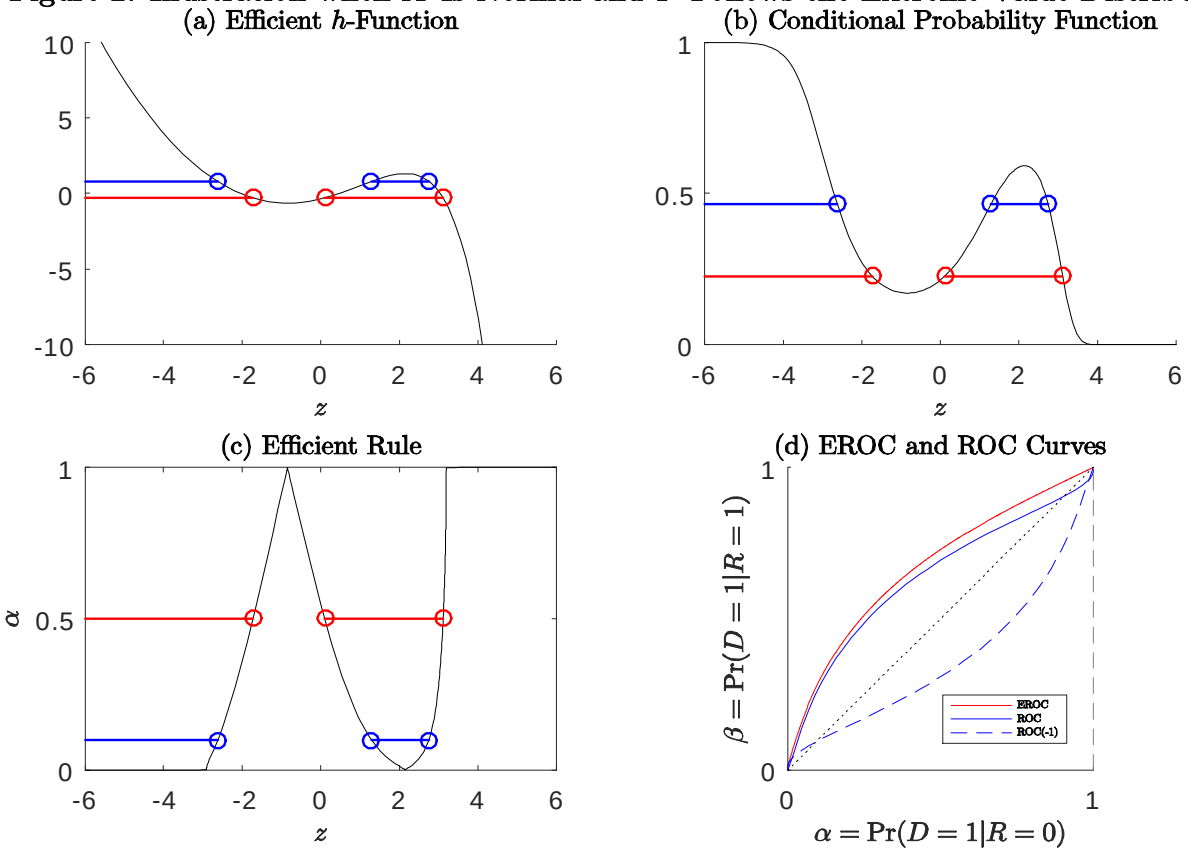
Vapnik, V. N., and A. Y. Chervonenkis (1971) "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability and Its Applications*, 16, 264–280.

Figure 1: Density Functions of X and Y



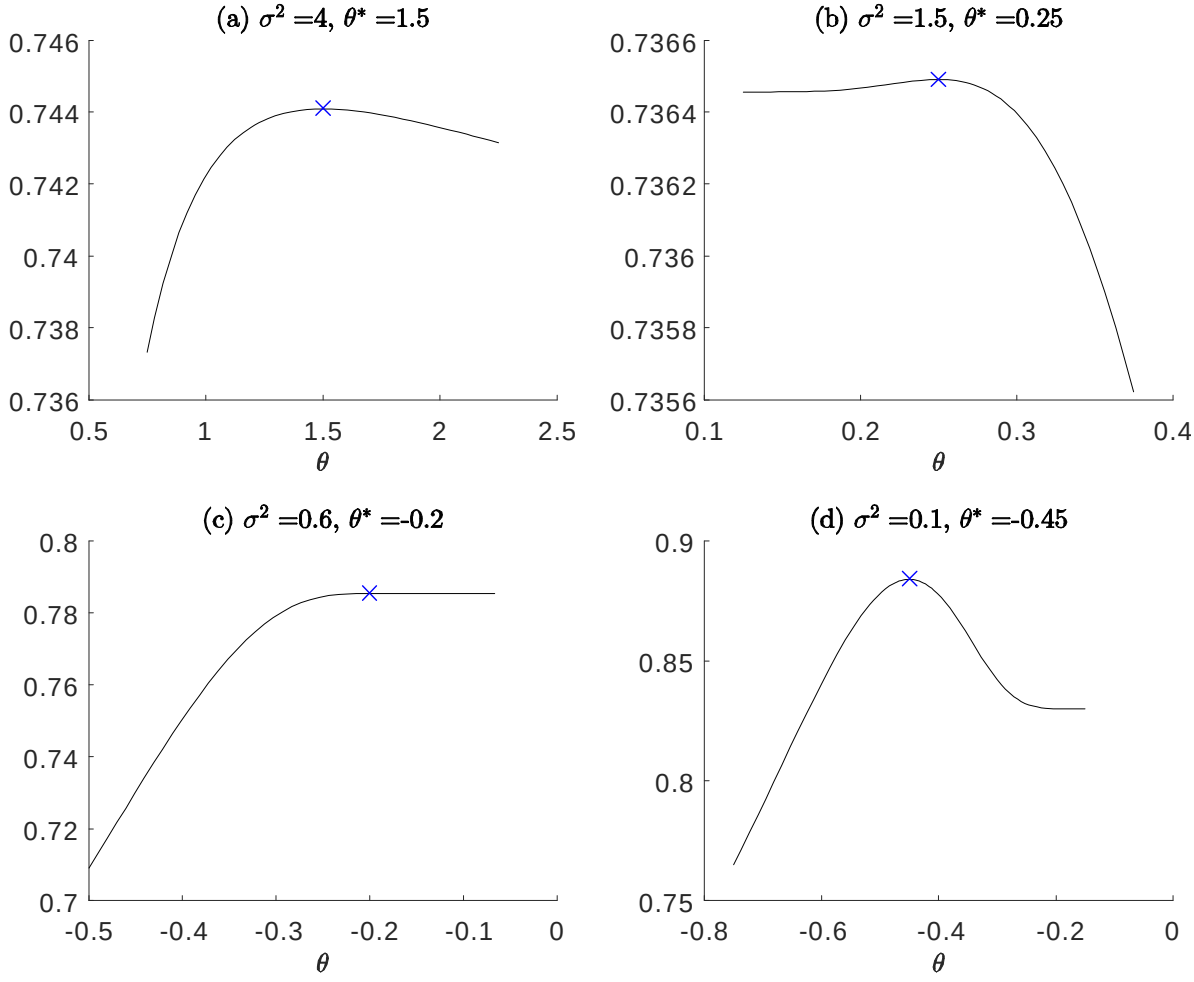
Notes: f and g, respectively, refers to the density function of standard normal and the extreme value distribution with location parameter and scale parameters equal to 1.

Figure 2: Illustration when X Is Normal and Y Follows the Extreme Value Distribution



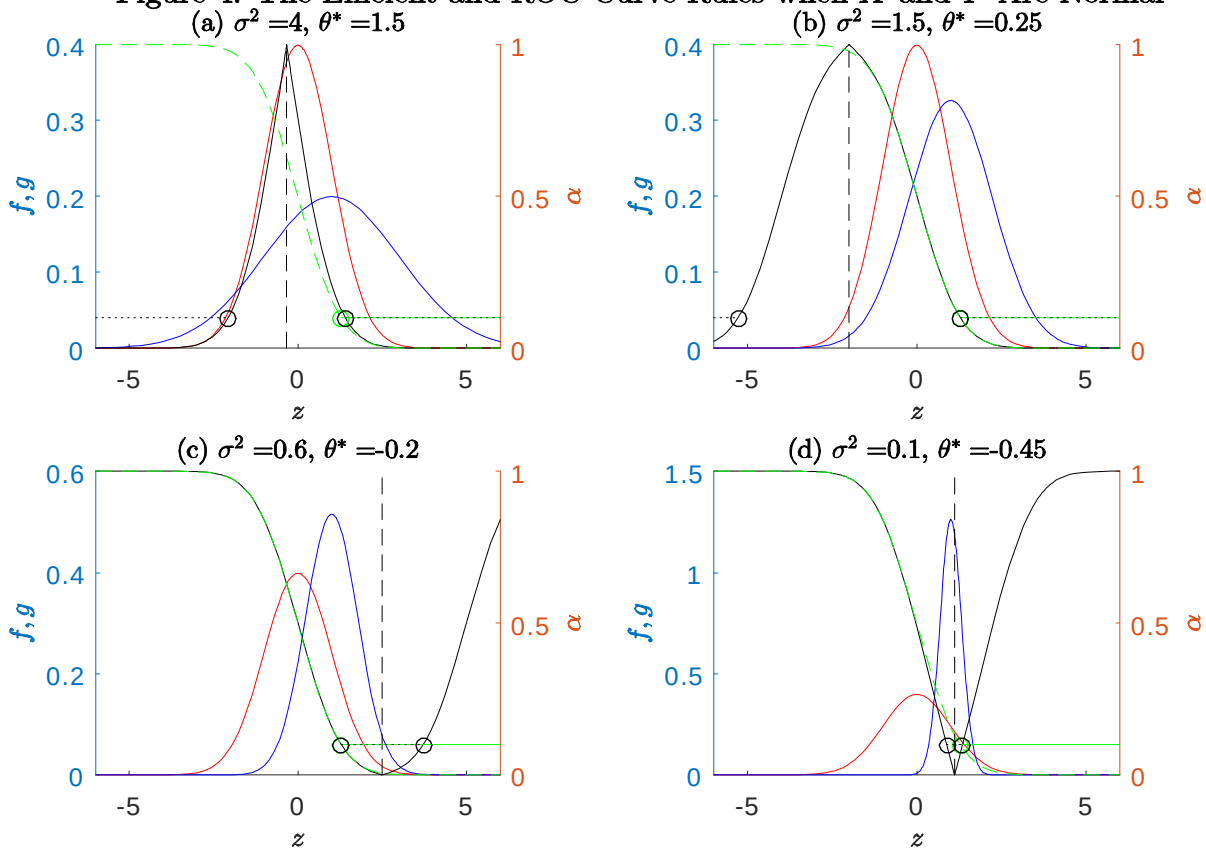
Notes: The figure applies in situations where there is a strictly monotone increasing function T such that $T(X)$ (i.e., $T(Z)|R = 0$) follows standard normal and $T(Y)$ (i.e., $T(Z)|R = 1$) follows the extreme value distribution with both the location parameter and the scale parameter equal to 1. Panel (a) plots the efficient h-function. Panel (b) plots the conditional probability function $\Pr(R = 1|Z = z)$ when $\Pr(R = 1) = 0.3$. Panel (c) plots the function $a^*(z) = 1 - F_{h^*}(h^*(z))$, where F_{h^*} is the distribution function of $h^*(X)$. The efficient prediction rule is given by $D^*_\alpha(Z) = I(a^*(Z) < \alpha)$, where $\alpha = \Pr(D^*_\alpha(X) = 1) \in [0, 1]$. In panels (a) through (c), the blue (the red) line indicates the set S_α such that, if $Z \in S_\alpha$, then $D^*_\alpha(Z) = 1$, $\alpha = 0.1$ ($\alpha = 0.5$).

Figure 3: GAUROC as a Function of θ when X is $N(0,1)$ and Y is $N(1,\sigma^2)$



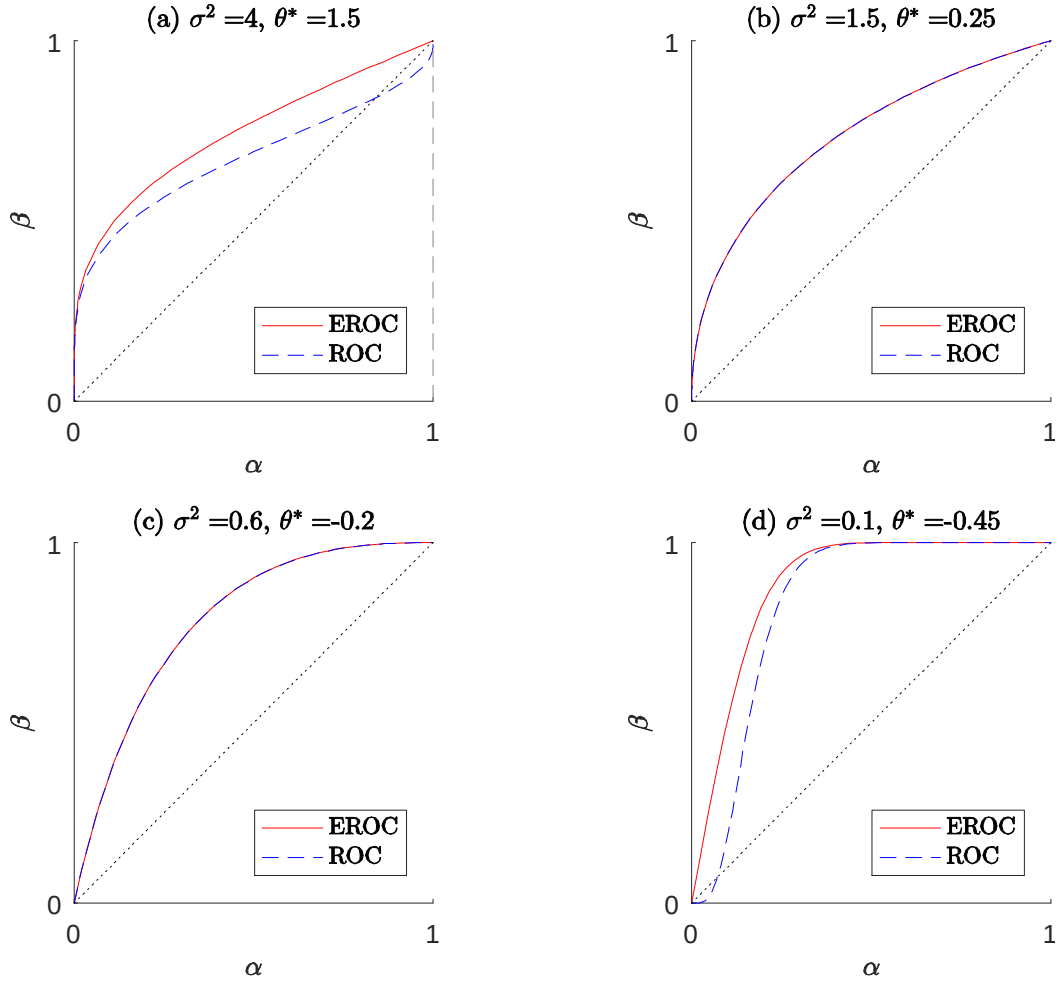
Notes: The figure plots GAUROC as a function θ for the rule $D(Z + \theta Z^2)$, when X (i.e., $Z|R = 0$) is standard normal and Y (i.e., $Z|R = 1$) is normal with mean 1 and variance σ^2. The panels differ by the value of σ^2 (or of $\theta^* = (\sigma^2 - 1)/2$). The blue cross indicates the efficient parameter θ^* in each case.

Figure 4: The Efficient and ROC Curve Rules when X and Y Are Normal



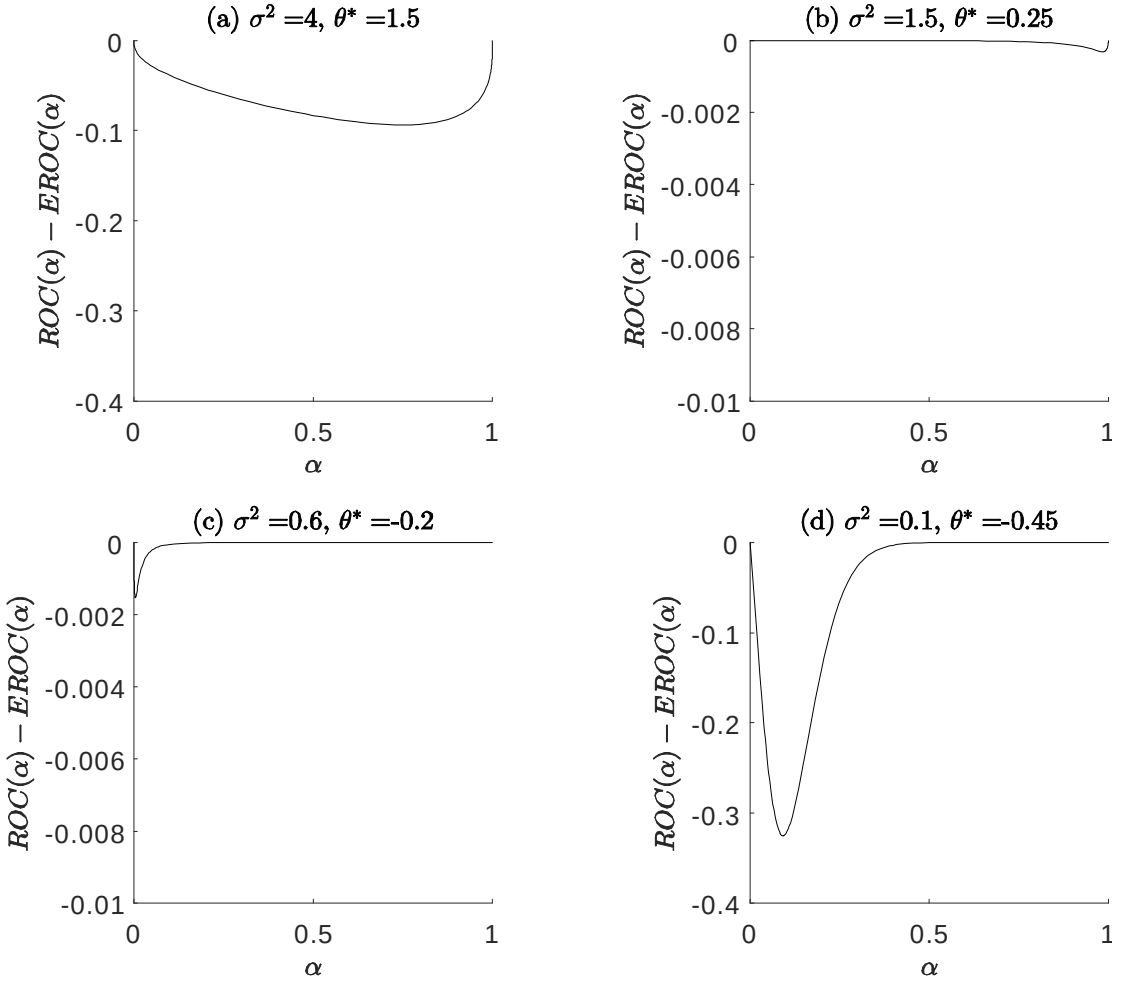
Notes: The figure plots the function $a(z) = 1 - F_h(h(z))$ for the efficient rule (the black solid line) and the ROC curve rule (the green dashed line), when X (i.e., $Z|R = 0$) is standard normal and Y (i.e., $Z|R = 1$) is normal with mean 1 and variance σ^2. The vertical black dashed line indicates the "turning point" of the efficient h-function, $h^*(z) = z + \theta^* z^2$. Recall that F_h is the distribution function of $h(X)$ and that the underlying rules can be expressed as $D(Z) = I(a(Z) < \alpha)$. The panels differ by the value of σ^2 (or of $\theta^* = (\sigma^2 - 1)/2$). The red (the blue) line depicts the density of X (of Y).

Figure 5: The EROC and ROC Curves when X and Y Are Normal



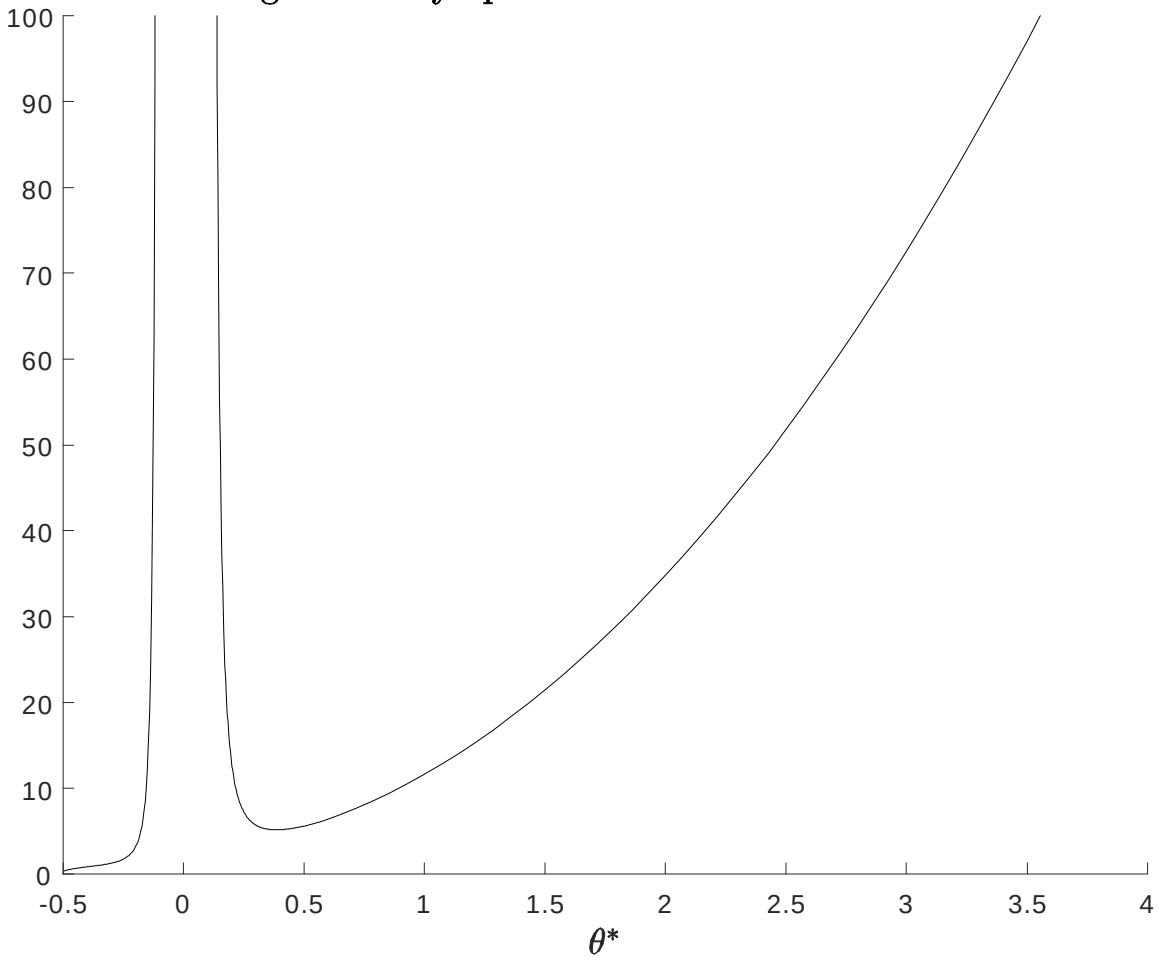
Notes: The figure plots the EROC and the ROC curves, when X (i.e., $Z|R=0$) is standard normal and Y (i.e., $Z|R=1$) is normal with mean 1 and variance σ^2.

Figure 6: Deviation of ROC about EROC



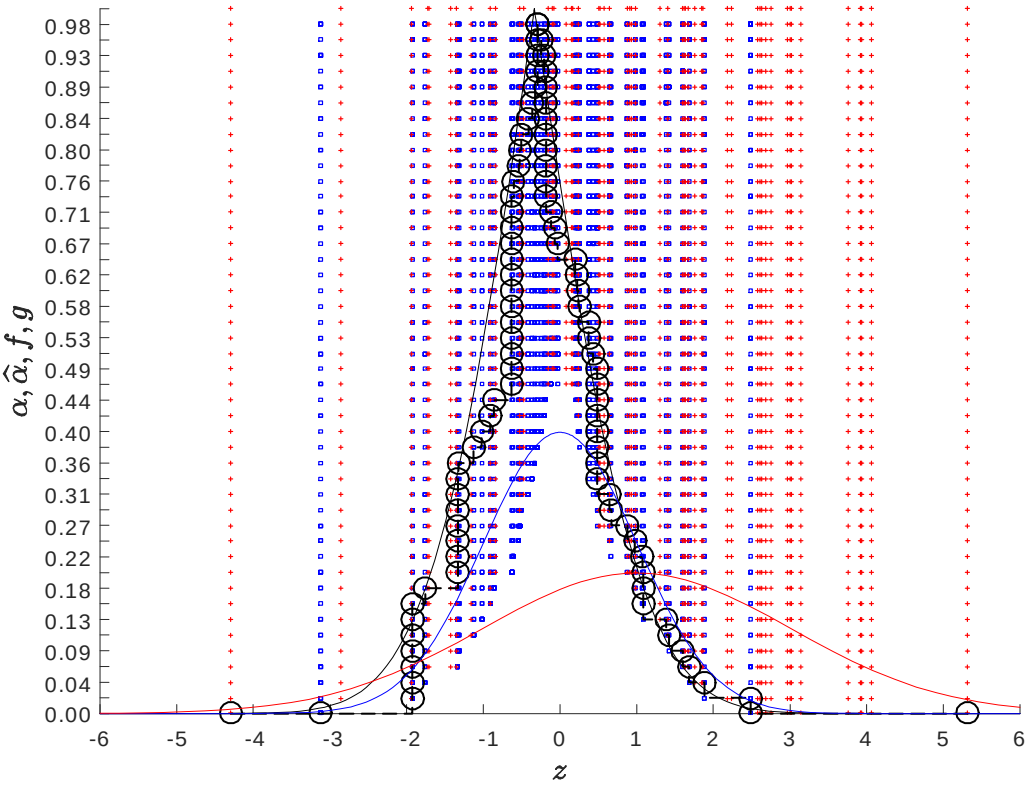
Notes: The deviations shown in the figure concern the ROC and the EROC curves in Figure 5.

Figure 7: Asymptotic Standard Deviation of $\widehat{\theta}$



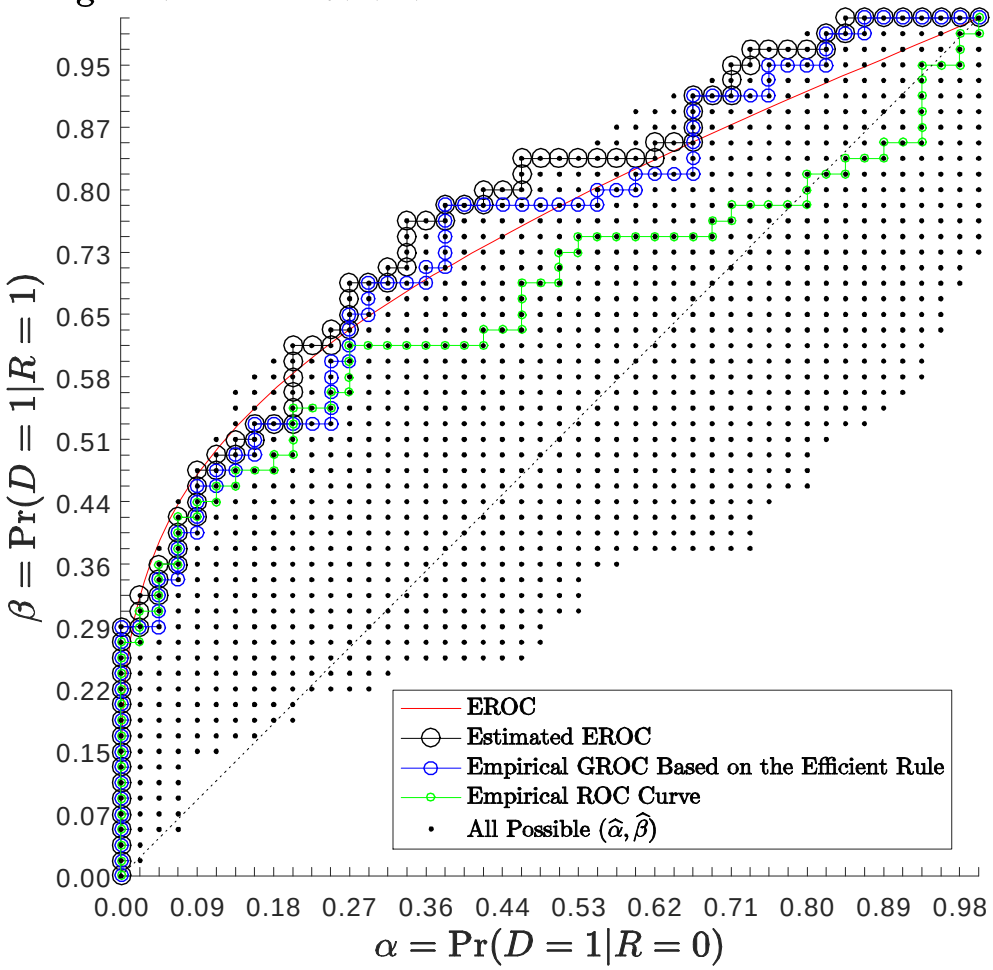
Notes: The figure plots the asymptotic standard deviation of $\widehat{\theta}$ as a function of $\theta^* = (\sigma^2 - 1)/2$, when X is $N(0, 1)$ and Y is $N(1, \sigma^2)$.

Figure 8: Efficient Rule and Its Nonparametric Estimate



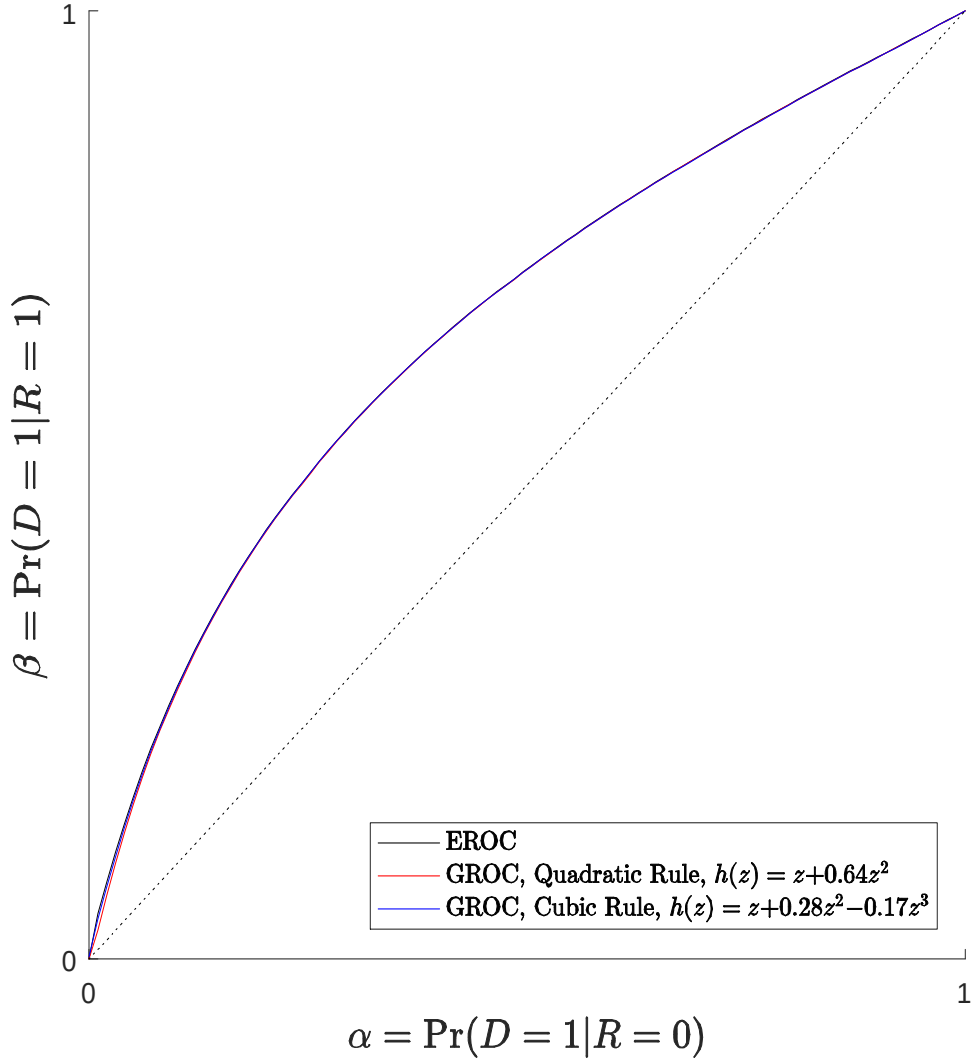
Notes: The figure illustrates the nonparametric estimate of the efficient rule based on a random sample of $n = 100$ observations on (R, Z), where $X = (Z|R = 0)$ is $N(0, 1)$, $Y = (Z|R = 1)$ is $N(1, 4)$, and $\Pr(R = 1) = 0.6$. The ticks at the vertical axes indicate the different estimates $\widehat{\alpha}$ of α that can be obtained from the sample by using the rule $D(a(Z) < \alpha)$, where $a(z) = 1 - F_h(h(z))$ is first increasing, then decreasing. The dotted line with circles is the nonparametric estimate of the efficient function $a^*(z) = 1 - F_{h^*}(h^*(z))$, where $h^*(z) = (z - \varrho)^2$ and F_{h^*} is the noncentral chi-distribution with noncentrality parameter $\varrho = -\frac{1}{3}$. The blue squares (the red crosses) refer to $n_X = 45$ ($n_Y = 55$) observations on X (Y), while the blue (the red) solid line is the population density of X (Y). At each tick mark, we plot only observations that can be used to obtain the corresponding estimate $\widehat{\alpha}$.

Figure 9: The EROC Curve and Its Non-Parametric Estimate



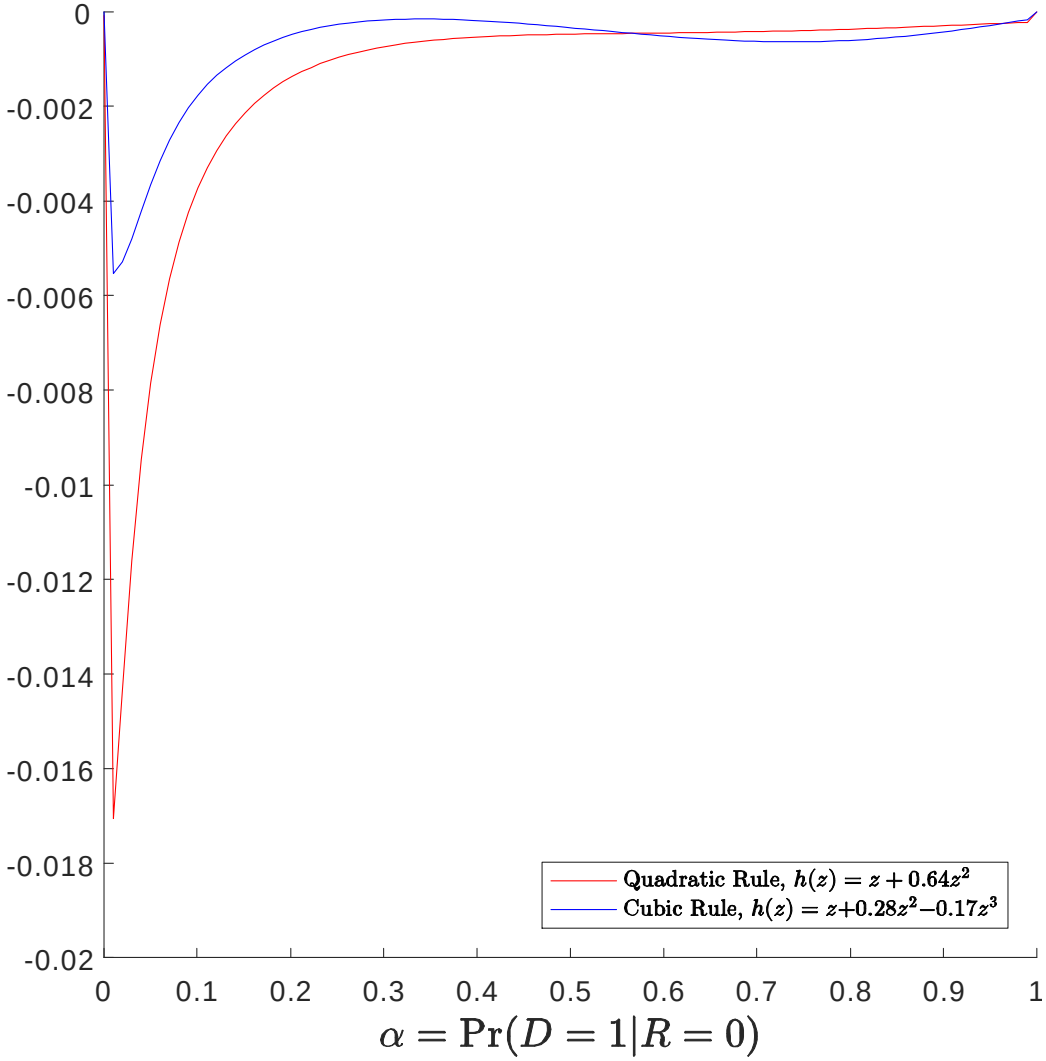
Notes: The black dots indicate all estimates $(\widehat{\alpha}, \widehat{\beta})$ that can be obtained from the random sample of Figure 8 by using either the rule in (27) or the ROC curve rule. The red solid line is the EROC curve based on the underlying true population (X standard normal, Y normal with mean 1 and variance 4). The dots marked by black circles and the associated black line constitute the empirical GROC curve based on the estimate of the efficient rule (the black dotted line with circles in Figure 8). The blue line with circles constitutes the empirical GROC curve based on the true efficient rule. The green line with circles is the empirical ROC curve.

Figure A1: Polynomial Approximations for an Efficient Rule



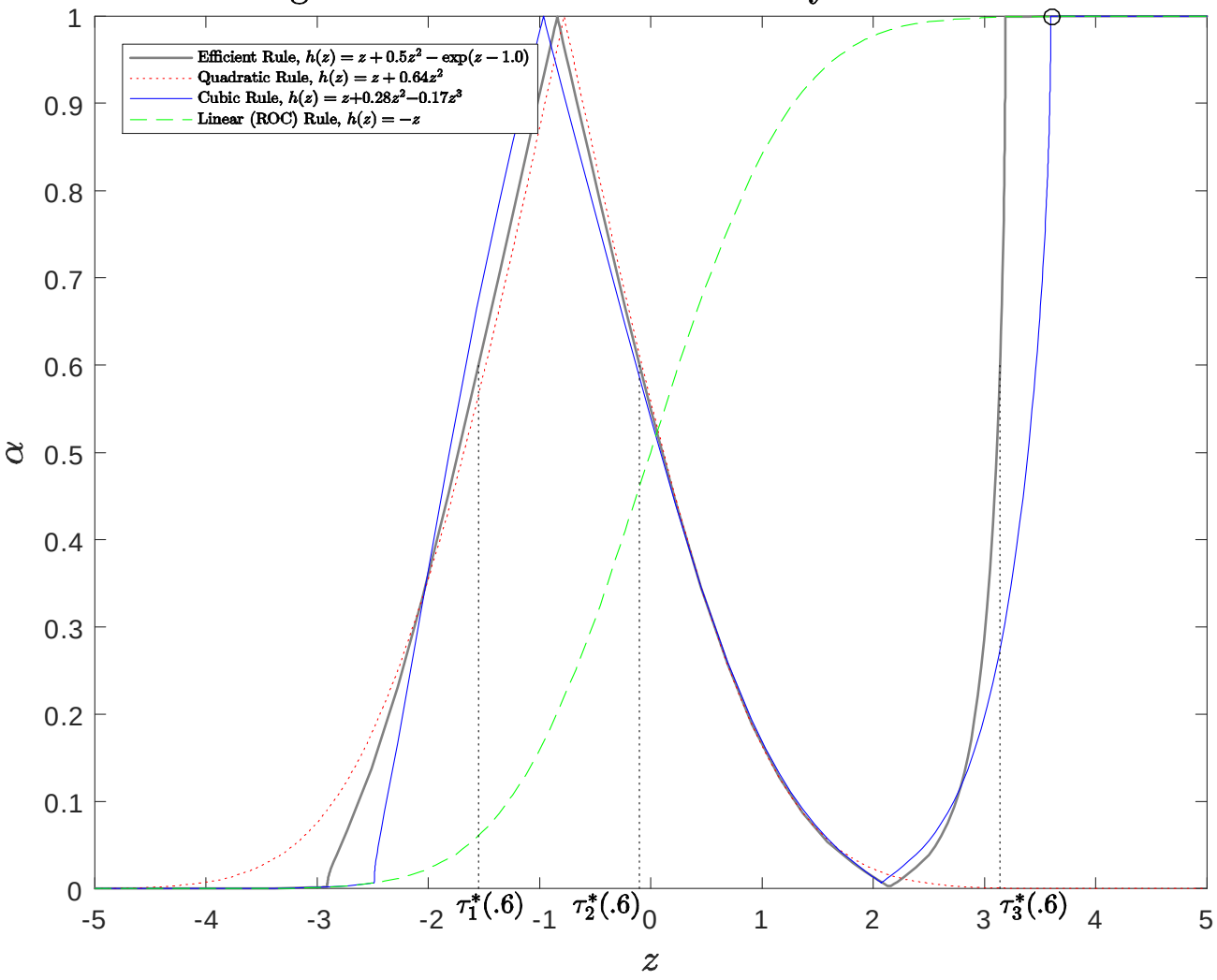
Notes: The figure shows the EROC curve along with the GROC curves for a "quadratic rule" $D(Z + 0.64Z^2)$ and a "cubic rule" $D(Z + 0.28Z^2 - 0.17Z^3)$, when X (i.e., $Z|R = 0$) follows standard normal and Y (i.e., $Z|R = 1$) follows the extreme value distribution with both the location parameter and the scale parameter equal to 1. The applied quadratic (cubic) rule yields the maximum GAUROC over all quadratic rules $D(Z + \theta Z^2)$ (cubic rules $D(Z + \theta_1 Z^2 + \theta_2 Z^3)$).

Figure A2: Deviation about EROC



Notes: The lines show the difference $GROC_h(\alpha) - EROC(\alpha)$ for the GAUROC maximizing quadratic and cubic rules when X (i.e., $Z|R = 0$) follows standard normal and Y (i.e., $Z|R = 1$) follows the extreme value distribution with both the location parameter and the scale parameter equal to 1.

Figure A3: Efficient Rule vs. Polynomial Rules



Notes: The curves show the function $a(z) = 1 - F_h(h(z))$ for the indicated h-functions when X (i.e., $Z|R = 0$) follows standard normal and Y (i.e., $Z|R = 1$) follows the extreme value distribution with both the location parameter and the scale parameter equal to 1. The function $a(z)$ describes the underlying decision rule as is explained in the text and in the notes of Figure 2 (panel (c)).

The **Aboa Centre for Economics (ACE)** is a joint initiative of the economics departments of the Turku School of Economics at the University of Turku and the School of Business and Economics at Åbo Akademi University. ACE was founded in 1998. The aim of the Centre is to coordinate research and education related to economics.

Contact information: Aboa Centre for Economics, Department of Economics, Rehtorinpellonkatu 3, FI-20500 Turku, Finland.

www.ace-economics.fi

ISSN 1796-3133