This is the working paper version of
Cook, J. A. (2017): "ROC curves and nonrandom data,"
Pattern Recognition Letters, 85(1), 35-41.

ROC curves and nonrandom data*

Jonathan Aaron Cook†

May 2016

Abstract

This paper shows that when a classifier is evaluated with nonrandom test data, ROC curves differ from the ROC curves that would be obtained with a random sample. To address this bias, this paper introduces a procedure for plotting ROC curves that are inferred from nonrandom test data. I provide simulations and an example with wine data to illustrate the procedure as well as the magnitude of bias that is found in standard ROC curves generated from nonrandom test data.

Keywords: ROC curves; Classifier evaluation; Sample-selection bias

1 Introduction

In many settings, data are collected in a nonrandom fashion. The decision to investigate insurance claims for fraud may be based on a predictive model. Investigating insurance claims is costly and it may be difficult to allocate resources to inspect a random sample of claims. Similarly, the Internal Revenue Service (IRS) uses a model that predicts tax-filing errors to select tax returns for audits. A recommender system may only show the user items that are predicted to be of interest. In these three examples, data are only collected for instances that are judged to be more likely to be positive cases.

This paper makes two contributions. This paper's first contribution is a characterization of the bias that results in receiver operating characteristic (ROC) curves when they are generated with nonrandom test data. Nonrandom test data can arise from using the classifier that we want to evaluate to select the test data or from using some other classifier to select the test data. This paper shows that the bias that arises in evaluating the classifier that was used to select the test data is similar to the well-known bias for regression with truncation on the dependent variable. When we only observe cases which are scored sufficiently high by the classifier, there is a type of attenuation bias for ROC curves. This paper also shows that the ROC curves are pushed outward for a classifier with low correlation to the classifier that was used to select the test data. This bias that arises when another classifier selected the test data is related to the bias for linear regression with incidental truncation.

This paper's second contribution is a procedure to create ROC curves that provide a consistent estimate of the ROC curve that would be obtained with random test data. This procedure infers the predictive power of the classifier based on available data and plots the implied ROC curve. The derived ROC curves are based on econometric work on bivariate probit analysis (e.g. Van de Ven and Van Pragg (1981) and Poirier (1980)). A key difference between this paper and

*This paper is under consideration at *Pattern Recognition Letters*

†Public Company Accounting Oversight Board. The PCAOB, as a matter of policy disclaims responsibility for any private publication or statement by any of its Economic Research Fellows and employees. The views expressed in this paper are the views of the author and do not necessarily reflect the views of the Board, individual Board members, or staff of the PCAOB. Email: JACook@uci.edu

		truth	
		positive	**negative**
prediction	**positive**	True Positives (TP)	False Positives (FP)
	negative	False Negatives (FN)	True Negatives (TN)
	total	Positives (P)	Negatives (N)

Table 1: Confusion matrix.

prior work on selection problems is that the problems considered by this paper are not regression equations. Section 4.2 discusses instances for which ROC curves are biased, but the parameters of a regression equation would not be. This paper makes distributional assumptions that lead to maximum likelihood problems that are similar to those encountered in estimating regression equations with truncation or incidental truncation. Under the distributional assumptions used in this paper, a classifier's expected ROC curve is determined by two parameters. The first parameter determines how many positive cases there are in the population. The second parameter is the correlation of the classifier's output with the true latent propensity to be a positive case.

This paper's procedure is related to the Dorfman-Alf (1969) procedure for estimating parameters of fitted ROC curves, which also uses maximum likelihood estimates under parametric assumptions. The Dorfman-Alf procedure does not correct for selecting test data with a classifier.

This paper contributes to the literature on evaluating classifiers. Recent works have shown the connections between ROC curves and precision-recall curves (Davis and Goadrich 2006) and cost curves (Hernández-Orallo et al. 2013). Other work on the properties of evaluation metrics for classifiers include Wang et al. (2013), who show that normalized discounted cumulative gains (NDCG) can consistently distinguish classifiers, and Moffat (2013), who provides properties of evaluation metrics. There does not appear to be any existing work on evaluating classifiers with nonrandom data.

This paper does not consider the problem of creating a classifier with nonrandom data. To create classifiers with nonrandom training data, the econometric literature has built on the sample-selection correction regression of Heckman (1976, 1979) (see Van de Ven and Van Pragg (1981) for a binary classifier). The credit-scoring literature has introduced *reject inference*, which incorporates information from unselected items, to improve classifier performance (see, for example, Crook and Banasik 2004).

In the next section, I derive the bias in ROC curves when the classifier being evaluated was used to select the test data. Section 3 derives a ROC curve that is consistent with nonrandom test data. I consider two cases of nonrandom test data:

(i) Using the classifier that we want to evaluate to select the test data, and

(ii) Using an unknown classifier to select the test data.

Monte Carlo simulations and an example with wine quality data are presented in Sections 4 and 5 to illustrate this procedure as well as the bias found in standard ROC curves. Section 6 concludes.

2 Classifiers and ROC curves

A *classifier* maps instances to predicted classes. This paper focuses on *binary classifiers*, which map to two classes (e.g., positive and negative). While some classifiers map directly to predicted classes, this paper focuses on classifiers that produce a continuous output. Given the classifier's output and a threshold, we classify all instances above the threshold as positive and all instances below the threshold as negative.

The confusion matrix in Figure 1 defines true positives (TP), true negatives (TN), positives (P), and negatives (N). We define sensitivity and specificity as

$$\text{Sensitivity} = \frac{TP}{P}, \quad \text{and} \tag{1}$$

$$\text{Specificity} = \frac{TN}{N}. \tag{2}$$

ROC curves, which plot sensitivity as a function of specificity for all possible thresholds, illustrate a classifier's trade-off between true positives and false negatives. A higher value of sensitivity for a given value of specificity indicates better performance. The area under the ROC curve (AUC) is a commonly used metric for evaluating a classifier's performance (as described by Bradley (1997)). If the classifier's output has no connection to the true class, the expected AUC would be .5. An excellent introduction to ROC curves is provided by Fawcett (2006).

2.1 Evaluating a classifier that was used to select the test data

This section shows that ROC curves are biased downward for the classifier that was used to select the test data. Let us denote the continuous output of classifier $\mathcal{A}$ for each instance i as a_i. I assume that there is some unobserved propensity to be a positive case and denote this propensity as p_i for each instance i. The true classification of each instance is

$$\text{outcome}_i = \begin{cases} \text{positive} & \text{if } p_i \geq p^* \\ \text{negative} & \text{otherwise} \end{cases}, \tag{3}$$

where p^* is the threshold for an instance to be a positive case. A value of $p^* = 0$ indicates that half of the observations are positive cases. The class skew increases with the absolute value of p^*. Throughout this paper, I treat both p_i and a_i as (possibly correlated) random variables. The modeler never observes p_i, only outcome_i. For a given threshold c, we can give probabilistic definitions of sensitivity and specificity:

$$\text{Sensitivity} = \text{Prob}(a_i > c \,|\, p_i > p^*), \quad \text{and} \tag{4}$$

$$\text{Specificity} = \text{Prob}(a_i < c \,|\, p_i < p^*). \tag{5}$$

The values in Equations (1) and (2) provide sample estimates of these probabilities.

This section considers the simplest form of choosing test data based on the classifier's output: choosing all instances with a value of a_i above s (for some constant s). We denote sensitivity and specificity conditional on selection as

$$\text{Sensitivity} \,|\, \text{Selection} = \text{Prob}(a_i > c \,|\, p_i > p^*, a_i > s), \quad \text{and} \tag{6}$$

$$\text{Specificity} \,|\, \text{Selection} = \text{Prob}(a_i < c \,|\, p_i < p^*, a_i > s). \tag{7}$$

When data are chosen based on the classifier's output, the estimates in Equations (1) and (2) provide an estimate of the values in Equations (6) and (7) instead of the values in Equations (4) and (5).

The following lemma will aid in proving our results regarding the bias in standard ROC curves for nonrandom test data.

Lemma 1 *For a fixed value of c, conditioning on selection*

(i) Increases sensitivity, i.e.

$$Sensitivity < Sensitivity \,|\, Selection \quad for\ all\ -\infty < s < c \quad , and$$

(ii) Decreases specificity, i.e.

$$Specificity > Specificity \,|\, Selection \quad for\ all\ -\infty < s < c.$$

All proofs are provided in the appendix. For a given cutoff level, selection moves sensitivity and specificity in opposite directions. The intuition for this result is that, as we focus on instances that our classifier considers more likely to be positive cases, we will have more positive cases in our test data. Sensitivity, which is conditional on the number of positive cases, is biased downward as the relative prevalence of positive cases increases. Similarly, specificity is biased upward as the relative number of negative cases decreases.

The ROC curve plots sensitivity as a function of specificity:

$$\begin{aligned} \text{Sensitivity}(\text{ Specificity}) = &\text{Prob}(a_i > c \,|\, p_i > p^*), \quad \text{where } c \text{ satisfies} \\ &\text{Specificity} = \text{Prob}(a_i < c \,|\, p_i < p^*). \end{aligned} \tag{8}$$

Up to this point, we have not made any distributional assumptions. To derive analytical results about the effect of selection on ROC curves, it is useful to assume that p_i and a_i come from a bivariate normal distribution:

$$\begin{pmatrix} p_i \\ a_i \end{pmatrix} \sim N_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{ap} \\ \rho_{ap} & 1 \end{bmatrix} \right).$$

The multivariate normal distribution is chosen because of the relative ease of working with conditional distributions. Given that the scale of the unobserved risk is arbitrary, I define the mean and variance of p_i to be zero and one. This is only done for notational simplicity and p_i can be redefined such that is has mean zero and variance one. The assumption that the classifier's output is normally distributed is an easily testable assumption.

We are now ready to state the main result of this section.

Proposition 2 *When test data are selected based on the classifier that we want to evaluate, sensitivity is lower for all values of specificity between zero and one.*

The assumed bivariate distribution is a sufficient but not necessary condition for Proposition 2. The downward bias in the ROC curve is created by truncating the distribution of the classifier's output. Truncation causes an attenuation bias in perceived correlation between the classifier's output and the latent propensity to be a positive case. This attenuation bias causes the ROC curve to cave in.

2.2 Evaluating a classifier with test data selected by an another classifier

I now consider the case of using another classifier, with output denoted as b, to select the test data. This paper focuses on situations in which b is not observed. Appendix B explores the case when b is observed. We assume that each instance of b can be written as

$$b_i = \delta\, X_i + \gamma\, a_i + \varepsilon_i,$$

where X_i is a vector of features for case i and ε_i is a standard normal random variable. The parameter δ is a vector of coefficients and γ indicates the degree to which the classifier's output

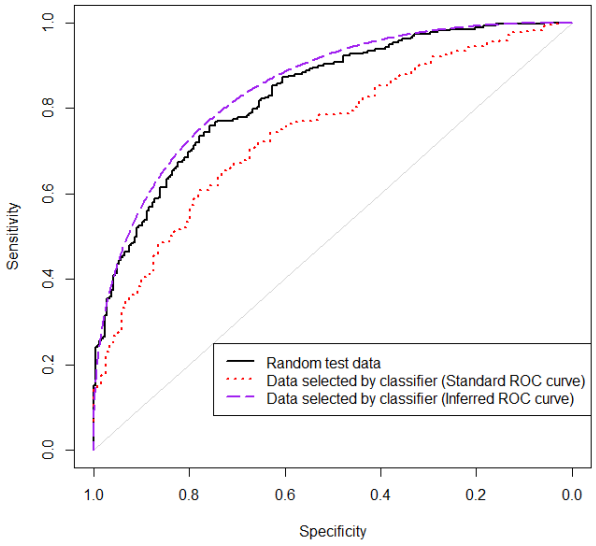

Figure 1: ROC curves for a classifier that was used to select the test data. The simulation above uses $\rho_{ap} = .7$ and $p^* = 0$. For data that was selected by the classifier, 1,000 instances are drawn from the bivariate normal distribution and the 500 draws with the greatest value of a are chosen.

was incorporated into the selection process. I assume that ε is mean independent of X and α, i.e. $E(\varepsilon|X, \alpha) = 0$. This assumption allows for estimation of δ and γ by a probit regression.

The selection rule is

$$\begin{cases} \text{Selected} & \text{if } \delta\, X_i + \gamma\, a_i + \varepsilon_i > s \\ \text{Not selected} & \text{otherwise} \end{cases} . \tag{9}$$

When $\delta = 0$, $\gamma = 1$, and $\text{Var}(\varepsilon) = 0$, this selection rule reduces to the case explored in Section 2.1. A positive correlation between ε and p indicates that information that used to select the test data, which is not included in a, is predictive of positive cases.

3 ROC curves for nonrandom test data

This paper's procedure for creating ROC curves that are robust to sample selection is to infer the predictive power of the classifier (taking truncation into consideration) then draw the ROC curve that is implied by our distributional assumptions. The proposed procedure has the following three steps.

Step 1 Subtract the mean and divide by the standard deviation to standardize the classifier's output. The mean and standard deviation should be based on all of the data, not only on the selected instances.

Step 2 Estimate p^* and the correlation between the classifier's output and the latent propensity to be a positive case, i.e. ρ_{ap}.

Step 3 Draw the ROC curve that is implied by our estimates in Step 2 and

$$\begin{aligned} \text{Sensitivity}(\ \text{Specificity}) =& \text{Prob}(a_i > c \,|\, p_i > p^*), \quad \text{where } c \text{ satisfies} \\ \text{Specificity} =& \text{Prob}(a_i < c \,|\, p_i < p^*). \end{aligned} \tag{10}$$

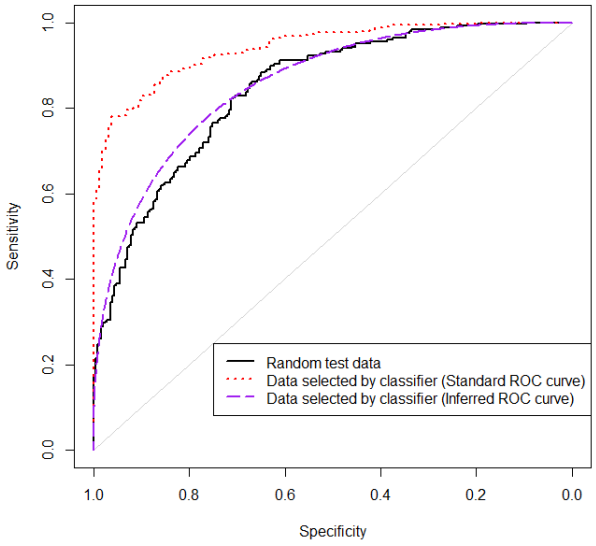

Figure 2: ROC curves with test data selected by another classifier. The simulation above use $\rho_{ap} = \rho_{\varepsilon p} = .7$, $\gamma = 0$, and $p^* = 0$. For data that was selected by the classifier, 1,000 instances are drawn from the bivariate normal distribution and the 500 draws with the greatest value of $\gamma\, a + \varepsilon$ are chosen.

To draw the ROC that implied by these estimates (denoted here as $\widehat{p^*}$ and $\widehat{\rho_{ap}}$), begin with a set of cutoffs with sufficiently large range (e.g., -4 to 4). For each cutoff $c \in [-4, 4]$, we find the corresponding value of sensitivity as

$$\text{Prob}(a_i > c \,|\, p_i > p^*) = \left[1 - \Phi(\widehat{p^*})\right]^{-1} \int_c^{\infty} \phi(a) \left[1 - \Phi\left(\left[\widehat{p^*} - \widehat{\rho_{ap}}\, a\right] / \sqrt{1 - \widehat{\rho_{ap}}^{\,2}}\right)\right] da \quad (11)$$

and specificity as

$$\text{Prob}(a_i < c \,|\, p_i < p^*) = \Phi(\widehat{p^*})^{-1} \int_{-\infty}^{c} \phi(a)\, \Phi\left(\left[\widehat{p^*} - \widehat{\rho_{ap}}\, a\right] / \sqrt{1 - \widehat{\rho_{ap}}^{\,2}}\right) da. \quad (12)$$

The ROC curve that we draw in Step 3 is a deterministic function of the maximum likelihood estimates from Step 2. By the functional invariance property of maximum likelihood estimates, we know that the ROC curve drawn in Step 3 is a consistent estimate of the expected ROC curve.

The remainder of this section derives the maximum likelihood estimates for p^* and the correlation between the classifier and the latent propensity to be a positive case. These maximum likelihood estimates, as well as Equations (11) and (12), are based on an assumption of multivariate normality. The example in Section 4 illustrates the performance of this procedure for a case when the classifier's output is not normally distributed.

3.1 Evaluating a classifier that was used to select the test data

After selecting on a, the likelihood function for the data can be expressed as

$$L = \prod_i \Phi(-[p^* - \rho_{ap}\, a_i]/\sqrt{1-\rho_{ap}^2})^{\mathbb{1}(\text{outcome}_i=\text{positive})} \times \Phi([p^* - \rho_{ap}\, a_i]/\sqrt{1-\rho_{ap}^2})^{\mathbb{1}(\text{outcome}_i=\text{negative})} \times \phi(a_i), \tag{13}$$

where $\mathbb{1}(\,)$ is the indicator function. We can find the maximum likelihood estimates for ρ_{ap} and p^* from

$$\widehat{\rho_{ap}}, \widehat{p^*} = \arg\max_{\rho_{ap}, p^*} \sum_i [\mathbb{1}(\text{outcome}_i = \text{positive}] \times \ln[\Phi(-[p^* - \rho_{ap}\, a_i]/\sqrt{1-\rho_{ap}^2})] + \sum_i [\mathbb{1}(\text{outcome}_i = \text{negative})] \times \ln[\Phi([p^* - \rho_{ap}\, a_i]/\sqrt{1-\rho_{ap}^2})] \tag{14}$$

The maximum likelihood estimates only depend on the instances that were selected by the classifier. The non-selected cases do not provide any additional information. An additional benefit of this procedure is that it provides an estimate of p^*. We can use the marginal distribution of p_i to infer the percent of positive cases in the population.

3.2 Evaluating a classifier with test data selected by another classifier

Under the selection rule

$$\begin{cases} \text{Selected} & \text{if } \delta\, X_i + \gamma\, a_i + \varepsilon_i > s \\ \text{Not selected} & \text{otherwise} \end{cases},$$

the likelihood function is

$$L = \prod_i \Phi_2(\delta\, X_i + \gamma\, a_i - s,\, -(p^* - a_i\, \rho_{ap})/\sqrt{1-\rho_{ap}^2}\,;\, \rho_{\varepsilon p})^{\mathbb{1}(\text{outcome}_i=\text{positive})} \times \Phi_2(\delta\, X_i + \gamma\, a_i - s,\, (p^* - a_i\, \rho_{ap})/\sqrt{1-\rho_{ap}^2}\,;\, -\rho_{\varepsilon p})^{\mathbb{1}(\text{outcome}_i=\text{negative})} \times \Phi(-[\delta\, X_i + \gamma\, a_i - s])^{\mathbb{1}(\text{outcome}_i=\text{NA})}. \tag{15}$$

This is a reparameterization of the likelihood derived by Van de Ven and Van Pragg (1981). The parameters δ, γ, s, ρ_{ap}, and $\rho_{\varepsilon p}$ can be estimated by maximizing the likelihood function in Equation (15).

4 Simulation

This section reports the results of simulation exercises for both of the procedures presented in the previous section. The purpose of these simulations is to illustrate the performance of inferred ROC curves as well as the bias that arises in standard ROC curves.

4.1 Evaluating a classifier that was used to select the test data

I first simulate the ROC curve that is obtained with random test data. For each Monte Carlo run, I draw 500 observations from the distribution

$$\begin{pmatrix} p_i \\ a_i \end{pmatrix} \sim N_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{ap} \\ \rho_{ap} & 1 \end{bmatrix}\right)$$

Evaluating a classifier that was used to select the test data

	$\rho_{ap} = .2$	$\rho_{ap} = .5$	$\rho_{ap} = .7$
AUC for ROC curves	.590	.730	.830
with a random sample	(.026)	(.022)	(.018)
AUC for standard ROC curves	.553	.643	.719
with data selected by a classifier	(.027)	(.025)	(.025)
AUC for inferred ROC curves	.591	.730	.830
with data selected by a classifier	(.041)	(.034)	(.027)
Portion positive cases in	.564	.666	.746
nonrandom test data	(.022)	(.022)	(.021)

Table 2: Results based on 10,000 simulations. Each simulation is based on a sample of 500 draws. Mean values across the simulations are presented with standard deviations in parenthesis. The parameter p^* is set to zero for all simulations so the portion of positive cases in the unbiased case is .5.

and define the outcome as

$$\text{outcome}_i = \begin{cases} \text{positive} & \text{if } p_i \geq p^* \\ \text{negative} & \text{otherwise} \end{cases} .$$

I then calculate the AUC for $\mathcal{A}$. This serves as an estimate of the unbiased AUC.

Next, I simulate the ROC curve that results with selection and with the inferred ROC curve. I draw 1,000 observations from this distribution, sort them by the values of a and keep the top 500. This is done so that there is a high degree of selectivity, but the number of observations used to generate the ROC curve is the same for both the random and nonrandom cases. I then use the 500 nonrandom observations to estimate ρ_{ap} and p^* based on Equation (14). These values are used to generate the ROC curve based on

$$\begin{gathered} \text{Sensitivity}(\text{ Specificity}) = \text{Prob}(a_i > c \,|\, p_i > \widehat{p^*}), \quad \text{where } c \text{ satisfies} \\ \text{Specificity} = \text{Prob}(a_i < c \,|\, p_i < \widehat{p^*}). \end{gathered}$$

I use a value of $p^* = 0$ for all simulations, but vary the value of ρ_{ap} across simulations. ROC curves for one simulation are presented in Figure 1. The biased ROC is caved in version of the ROC that is obtained with random data. The inferred ROC curve presents a smoothed out version of the unbiased ROC curve.

I use 100,000 Monte Carlo runs for each value of ρ_{ap}. These results are provided in Table 2. For better classifiers, which have larger values of ρ_{ap}, the bias of the AUC of standard ROC curves is larger. When $\rho_{ap} = .7$, the AUC for standard ROC when the data are selected by the classifier is 13% less than the AUC with a random sample (.719 compared with .830). The bias when $\rho_{ap} = .2$ is only 6% (.553 compared with .590). In each case, the average area under the inferred ROC curves matches the average AUC that is obtained with a random sample.

4.2 Evaluating a classifier with test data selected by another classifier

For each Monte Carlo run, I draw 1,000 observations from the distribution

$$\begin{pmatrix} p_i \\ a_i \\ \varepsilon_i \end{pmatrix} \sim N_3 \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{ap} & \rho_{\varepsilon p} \\ \rho_{ap} & 1 & 0 \\ \rho_{\varepsilon p} & 0 & 1 \end{bmatrix} \right)$$

Evaluating a classifier with test data selected by another classifier			
	$\rho_{ap} = .2$	$\rho_{ap} = .5$	$\rho_{ap} = .7$
AUC for ROC curves	.590	.730	.830
with a random sample	(.026)	(.022)	(.018)
For no correlation between classifiers and data selected by a highly predictive classifier, i.e. $\gamma = 0$, $\rho_{\varepsilon p} = .7$			
AUC for standard ROC curves	.619	.804	.936
with data selected by a classifier	(.028)	(.022)	(.011)
AUC for inferred ROC curves	.591	.734	.854
with data selected by a classifier	(.031)	(.058)	(.042)
For .5 **correlation between classifiers and data selected by a highly predictive classifier, i.e.** $\gamma = 1$, $\rho_{\varepsilon p} = .7$			
AUC for standard ROC curves	.578	.541	.683
with data selected by a classifier	(.028)	(.032)	(.059)
AUC for inferred ROC curves	.579	.708	.779
with data selected by a classifier	(.084)	(.084)	(.095)
For .5 **correlation between classifiers and data selected by a less predictive classifier, i.e.** $\gamma = 1$, $\rho_{\varepsilon p} = 0$			
AUC for standard ROC curves	.575	.698	.796
with data selected by a classifier	(.026)	(.024)	(.021)
AUC for inferred ROC curves	.591	.730	.832
with data selected by a classifier	(.093)	(.077)	(.057)

Table 3: Results based on 10,000 simulations. Each simulation is based on a sample of 500 draws. Mean values are presented with standard deviations in parenthesis. The parameter p^* is set to zero for all simulations.

and define the outcome as before. I then sort the values by $(\gamma a_i + \varepsilon_i)$ and keep the 500 largest. This is equivalent to setting $\delta = 0$ in Equation (9).

Across simulations, I vary the correlation between the classifiers and the correlation of the classifier that was used to select the test data with the latent propensity to be a positive case. The correlation between the classifiers is

$$\text{Cor}(a_i, b_i) = \frac{\gamma}{1 + \gamma^2}$$

and the correlation between the classifier that was used to select the test data and the latent propensity to be a positive case is

$$\text{Cor}(p_i, b_i) = \frac{\gamma\, \rho_{ap} + \rho_{\varepsilon p}}{1 + \gamma^2}.$$

When there is no correlation between the classifiers ($\gamma = 0$), there is an upward bias in the AUC for standard ROC curves. This is related to the tendency of ROC curves to be "overly optimistic" when the data is skewed (Davis and Goadrich 2006, p. 233). The bias is largest (13%) when $\rho_{ap} = \rho_{\varepsilon p} = .7$. The cases for which $\gamma = 0$ illustrate another difference between the problem considered by this paper and the econometric literature on sample-selection bias. For a regression, when there is no correlation between the selection rule and the regressors in the equation of interest, there is no bias for ordinary least squares regression.

When there is a .5 correlation between the classifiers (i.e. $\gamma = 1$), the ROC curve is biased downward. For small positive values of γ (results not reported), there is an upward bias in ROC curves. As the correlation between the classifiers increases, the bias becomes more similar to the truncation bias explored in Section 4.1. The cases for which $\rho_{\varepsilon p} = 0$ illustrate another difference between the problem considered by this paper and the econometric literature on sample-selection bias. For a regression, when there is no correlation between the stochastic element in the selection equation and the stochastic element in the outcome equation, there is no bias. By contrast, the bottom panel of Table 3 shows that there is a downward bias for ROC curves in this situation.

For $\gamma = 1$ and $\rho_{\varepsilon p} = .7$, there is a noticeable difference between the areas under the inferred ROC curve and the AUCs that are obtained with random test data. While these values are closer to the AUCs obtained with random test data than standard ROC curves with nonrandom test data, the standard deviations of the AUCs is much larger for inferred ROC curves. In results not shown, the differences between area under the inferred ROC curve and the AUC obtained with random test data are decreasing in the size of the sample. Appendix B shows that, when the classifier that selected the test data is observed, the average and standard deviation of AUCs from inferred ROC curves are equal to those obtained with random test data.

5 An example with wine-quality data

To provide a demonstration of this procedure with non-simulated data, I use data on white wine quality from Cortez et al. (2009).[1] This dataset contains eleven attributes for 4,898 white wines, including alcohol content, citric acid, and residual sugar. A detailed description of this data are provided by Cortez et al. For the measure of wine quality, each wine was evaluated by experts and given a score from zero to ten (with ten being the highest quality). Because we are interested in binary prediction, I define a wine with a score of six or higher as "good wine" and other wines as "not good wine."

I use all eleven attributes in a random forest classifier (based on Breiman (2001) and implemented in R using Liaw and Wiener's (2002) randomForest package) to predict (binary) wine

[1] These data are available at the University of California at Irvine's Machine Learning Repository, `https://archive.ics.uci.edu/ml/datasets.html`.

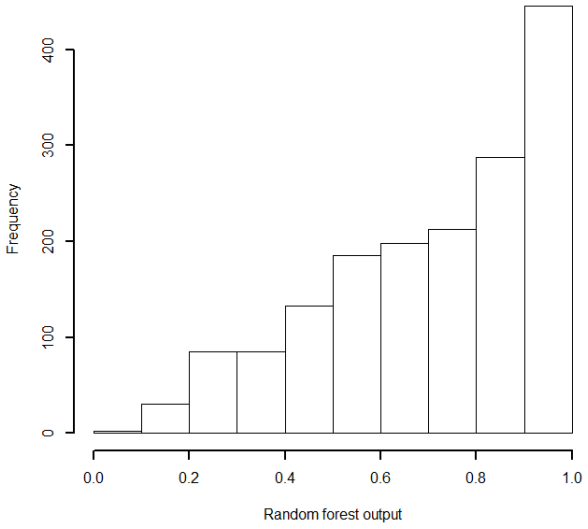

Figure 3: Histogram of the random forest's output. The mean, variance, and skew are .71, .05, and -.62, respectively.

quality. The random forest contains 1,000 trees and tries three attributes at each split. I use the first 2/3 of the observations (3,233 observations) as the training data and the remaining 1/3 (1,665 observations) as the test data.

I first find the ROC curve for the random forest classifier using the full set of test data. The area under the ROC curve is .83. Next, let us suppose that the wine experts do not have enough time to score all of the wine in the test data. Preferring to taste wine that is more likely to be good wine, the experts taste the half the test data that the random forest classifier predicted was most likely to be good wine. With only half of the test data available, the area under the ROC curve falls to .60.

I now perform the procedure described in Section 3 with the half of the test data predicted to be good wine. Figure 3 presents a histogram of the random forest's scores. The distribution is clearly non-Gaussian. This example provides some insight into the performance of this paper's proposed procedure when its assumptions are not met.

I standardize the random forest scores and maximize Equation (9) to find the estimates $\widehat{\rho_{ap}} = .64$ and $\widehat{p^*} = -.55$.

Figure 3 plots the ROC curves that are obtained with the full set of test data, the half of the test data that received a high score from the random forest, and the ROC curve based on our estimates of p^* and ρ_{ap}. The ROC curve based on our estimates of p^* and ρ_{ap} closely matches the ROC curve obtained with the full set of test data.

6 Discussion and Conclusion

For test data that is chosen by the classifier that we want to evaluate, ROC curves are biased downward. When test data was selected by another classifier, the direction of the bias in the ROC is not clear. This paper presents a procedure for creating ROC curves that provide a consistent estimate of the ROC curve that would be obtained with random test data.

The procedure introduced here relies on distributional assumptions. The example in Section

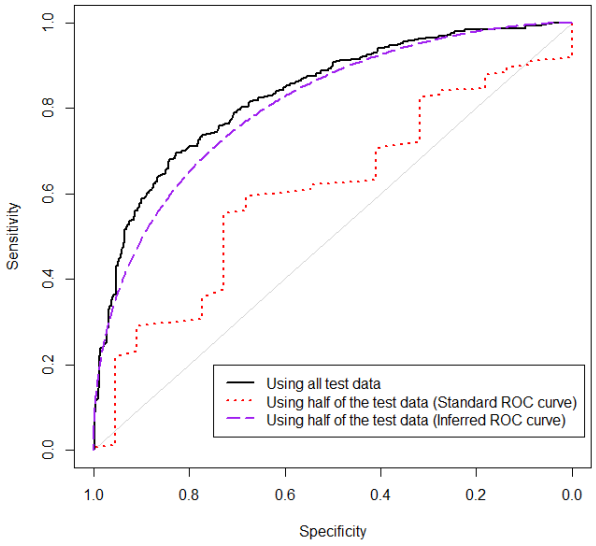

Figure 4: ROC curves for wine quality prediction, as described in Section 5. The area under the ROC curve that uses all of the test data is .83. The areas under the standard and inferred ROC curves are .60 and .81, respectively.

5 violates the assumed distribution for the classifier's output and the area under the inferred ROC curve is a still a close match to the area under the ROC curve that would be obtained with the full set of test data. If non-Gaussian distributions are preferred for classifiers' output, the multivariate distributions used in this paper could be written in terms of copulas, as has been done for sample-selection bias in a regression setting (as in Li and Rahman (2011)).

An advantage of these distributional assumptions is that they introduce a new parameter which measures the correlation between the classifier's output and latent propensity to be a positive case. Also, given that our inferred ROC curves are based on maximum likelihood estimates, it is possible to construct confidence bands for the curves.

Another advantage of these distributional assumptions is that the estimation of the parameter p^* leads an estimate of the percent of positive cases in the population. This parameter may be of interest to an organization like the IRS that could use $\widehat{p^*}$ to estimate the percent of tax returns that contain errors.

A Proofs

Proof of Lemma 1.
For (i):

We want to show that

$$\text{Prob}(a_i > c \,|\, p_i > p^*) < \text{Prob}(a_i > c \,|\, p_i > p^*, a_i > s).$$

We assume that $\text{Prob}(p_i > p^*)$ and $\text{Prob}(a_i > s)$ are both nonzero. If our selection rule s were negative enough, selection would have no impact on sensitivity:

$$\lim_{s \to -\infty} \text{Prob}(a_i > c \,|\, p_i > p^*, a_i > s) = \text{Prob}(a_i > c \,|\, p_i > p^*).$$

We will show that sensitivity is monotonically increasing in the selection rule s. We first rewrite specificity in terms of the pdf of a_i conditional on $(p_i > p^*)$ as

$$\text{Prob}(a_i > c \,|\, p_i > p^*, a_i > s) = \frac{\text{Prob}(a_i > c \,|\, p_i > p^*)}{\text{Prob}(a_i > s \,|\, p_i > p^*)} = \frac{\int_c^\infty f_{a|p>p^*}(a_i)da_i}{\int_s^\infty f_{a|p>p^*}(a_i)da_i},$$

where $f_{a|p>p^*}$ is pdf of a_i conditional on $(p_i > p^*)$. We take the derivative of specificity conditional on selection with respect to s through a straight-forward application of Leibniz rule:

$$\frac{d}{ds}\left(\frac{\int_c^\infty f_{a|p>p^*}(a_i)da_i}{\int_s^\infty f_{a|p>p^*}(a_i)da_i}\right) = f_{a|p>p^*}(s)\frac{\int_c^\infty f_{a|p>p^*}(a_i)da_i}{\left[\int_s^\infty f_{a|p>p^*}(a_i)da_i\right]^2} > 0.$$

For (ii):

We want to show that

$$\text{Prob}(a_i < c \,|\, p_i < p^*) > \text{Prob}(a_i < c \,|\, p_i < p^*, a_i > s).$$

We assume that $\text{Prob}(p_i < p^*)$ and $\text{Prob}(a_i > s)$ are both nonzero. As in part (i), we begin by noting that if our selection rule s were negative enough, selection would have no impact on specificity:

$$\lim_{s\to-\infty} \text{Prob}(a_i < c \,|\, p_i < p^*, a_i > s) = \text{Prob}(a_i < c \,|\, p_i < p^*).$$

We will show that specificity is monotonically decreasing in the selection rule s. We first rewrite specificity in terms of the pdf of a_i conditional on $(p_i < p^*)$ as

$$\text{Prob}(a_i < c \,|\, p_i < p^*, a_i > s) = \frac{\text{Prob}(s < a_i < c \,|\, p_i < p^*)}{\text{Prob}(a_i > s \,|\, p_i < p^*)} = \frac{\int_s^c f_{a|p<p^*}(a_i)da_i}{\int_s^\infty f_{a|p<p^*}(a_i)da_i},$$

where $f_{a|p<p^*}$ is pdf of a_i conditional on $(p_i < p^*)$. We take the derivative of specificity conditional on selection with respect to s through a straight-forward application of Leibniz rule:

$$\frac{d}{ds}\left(\frac{\int_s^c f_{a|p<p^*}(a_i)da_i}{\int_s^\infty f_{a|p<p^*}(a_i)da_i}\right) = -\frac{f_{a|p<p^*}(s)\left[\int_s^\infty f_{a|p<p^*}(a_i)da_i - \int_s^c f_{a|p<p^*}(a_i)da_i\right]}{\left[\int_s^\infty f_{a|p<p^*}(a_i)da_i\right]^2} < 0.$$

■

Proof of Proposition 2.

Here, I show that sensitivity for a given level of specificity is a decreasing function of s. Since this term approaches a point on the ROC curve as s approaches negative infinity, a monotonic decrease in s implies that any point on the ROC curve will be lower.

I define sensitivity for a given level of specificity and selection rule s as

$$\begin{aligned}\text{Sensitivity(}\ \text{Specificity, }\ s\ \text{)} =& \text{Prob}(a_i > c \,|\, p_i > p^*, a_i > s), \quad \text{where } c \text{ satisfies} \\ & \text{Specificity} = \text{Prob}(a_i < c \,|\, p_i < p^*, a_i > s),\end{aligned}$$

assuming that $\text{Prob}(p_i < p^* \,|\, a_i > s)$, $\text{Prob}(p_i > p^* \,|\, a_i > s)$, and $\text{Prob}(a_i > s)$ are all nonzero. For a fixed level of specificity, the effect of an increase in s on sensitivity is

$$\frac{d\,\text{Sensitivity}}{d\,s} = \underbrace{\frac{\partial\,\text{Sensitivity}}{\partial\,s}}_{\text{Direct effect of } s \text{ on sensitivity}} + \underbrace{\frac{\partial\,\text{Sensitivity}}{\partial\,c}\frac{d\,c}{d\,s}}_{\text{Indirect effect of changing } c}.$$

These terms are

$$\begin{aligned}
\frac{\partial\,\text{Sensitivity}}{\partial\, s} &= f_{a|p>p^*}(s)\frac{\int_c^\infty f_{a|p>p^*}(a_i)da_i}{\left[\int_s^\infty f_{a|p>p^*}(a_i)da_i\right]^2} > 0,\\
\frac{\partial\,\text{Sensitivity}}{\partial\, c} &= -\frac{f_{a|p>p^*}(c)}{\int_s^\infty f_{a|p>p^*}(a_i)da_i} < 0, \quad \text{and}\\
\frac{d\, c}{d\, s} &= -\frac{\partial\,\text{Prob}(a_i < c\,|\,p_i < p^*, a_i > s)/\partial\, c}{\partial\,\text{Prob}(a_i < c\,|\,p_i < p^*, a_i > s)/\partial\, s}\\
&= \frac{f_{a|p<p^*}(c)\left[\int_s^\infty f_{a|p<p^*}(a_i)da_i\right]}{f_{a|p<p^*}(s)\left[\int_c^\infty f_{a|p<p^*}(a_i)da_i\right]} > 0,
\end{aligned}$$

where the last term follows from the use of the implicit function theorem. After applying some high-school algebra, $d\,\text{Sensitivity}/d\, s$ can be written as

$$\begin{aligned}
\frac{d\,\text{Sensitivity}}{d\, s} =& \frac{f_{a|p>p^*}(s)f_{a|p<p^*}(s)\left[\int_c^\infty f_{a|p>p^*}(a_i)da_i\right]\left[\int_c^\infty f_{a|p<p^*}(a_i)da_i\right]}{f_{a|p<p^*}(s)\left[\int_c^\infty f_{a|p<p^*}(a_i)da_i\right]\left[\int_s^\infty f_{a|p>p^*}(a_i)da_i\right]^2}\\
&- \frac{f_{a|p>p^*}(c)f_{a|p<p^*}(c)\left[\int_s^\infty f_{a|p>p^*}(a_i)da_i\right]\left[\int_s^\infty f_{a|p<p^*}(a_i)da_i\right]}{f_{a|p<p^*}(s)\left[\int_c^\infty f_{a|p<p^*}(a_i)da_i\right]\left[\int_s^\infty f_{a|p>p^*}(a_i)da_i\right]^2}.
\end{aligned}$$

The denominator is clearly positive so we focus on the numerator. Given the bivariate distribution that we assumed, the condition for the numerator to be negative is

$$\begin{aligned}
&[\phi(s)]^2\,\Phi(-\rho_{ap}s/(1-\rho_{ap})^2)\Phi(\rho_{ap}s/(1-\rho_{ap})^2)\\
&\times\int_c^\infty \phi(a_i)\Phi(-a_i\rho_{ap}/(1-\rho_{ap}^2))da_i\int_c^\infty \phi(a_i)\Phi(a_i\rho_{ap}/(1-\rho_{ap}^2))da_i\\
&< [\phi(c)]^2\,\Phi(-\rho_{ap}c/(1-\rho_{ap})^2)\Phi(\rho_{ap}c/(1-\rho_{ap})^2)\\
&\times\int_s^\infty \phi(a_i)\Phi(-a_i\rho_{ap}/(1-\rho_{ap}^2))da_i\int_s^\infty \phi(a_i)\Phi(a_i\rho_{ap}/(1-\rho_{ap}^2))da_i.
\end{aligned}$$

This condition holds for $c > s$. It follows that

$$\frac{d\,\text{Sensitivity}}{d\, s} < 0,$$

which implies that, for a fixed level of specificity, sensitivity is monotonically decreasing in selectivity s.

■

B Evaluating a classifier with test data selected by an observed classifier

As in the main text, I consider the case of selection of test data based another classifier, b. Instance i is selected if $b_i > s$ and not selected otherwise. Unlike the main text, this section explores situations in which b is observed. As before, I allow for correlation between the output of classifiers $\mathcal{A}$ and $\mathcal{B}$, which could arise from using similar attributes to make predictions:

$$\begin{pmatrix} p_i \\ a_i \\ b_i \end{pmatrix} \sim N_3\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{ap} & \rho_{bp} \\ \rho_{ap} & 1 & \rho_{ab} \\ \rho_{bp} & \rho_{ab} & 1 \end{bmatrix}\right).$$

Evaluating a classifier with test data selected by an observed classifier			
	$\rho_{ap} = .2$	$\rho_{ap} = .5$	$\rho_{ap} = .7$
AUC for ROC curves	.590	.730	.830
with a random sample	(.026)	(.022)	(.018)
For no correlation between classifiers and data selected by a highly predictive classifier, i.e. $\rho_{ab} = 0$, $\rho_{bp} = .7$			
AUC for standard ROC curves	.619	.804	.936
with data selected by a classifier	(.029)	(.022)	(.011)
AUC for inferred ROC curves	.590	.730	.829
with data selected by a classifier	(.023)	(.021)	(.015)
For .5 **correlation between classifiers and data selected by a highly predictive classifier, i.e.** $\rho_{ab} = .5$, $\rho_{bp} = .7$			
AUC for standard ROC curves	.533	.666	.800
with data selected by a classifier	(.027)	(.027)	(.021)
AUC for inferred ROC curves	.590	.730	.829
with data selected by a classifier	(.025)	(.020)	(.015)
For .5 **correlation between classifiers and data selected by a less predictive classifier, i.e.** $\rho_{ab} = .5$, $\rho_{bp} = .2$			
AUC for standard ROC curves	.568	.722	.832
with data selected by a classifier	(.026)	(.023)	(.018)
AUC for inferred ROC curves	.591	.730	.830
with data selected by a classifier	(.028)	(.025)	(.020)
For .5 **correlation between classifiers and data selected by a nonpredictive classifier, i.e.** $\rho_{ab} = .5$, $\rho_{bp} = 0$			
AUC for standard ROC curves	.599	.752	.863
with data selected by a classifier	(.025)	(.022)	(.016)
AUC for inferred ROC curves	.591	.730	.830
with data selected by a classifier	(.029)	(.026)	(.021)

Table 4: Results based on 10,000 simulations. Each simulation is based on a sample of 500 draws. Mean values are presented with standard deviations in parenthesis. The parameter p^* is set to zero for all simulations.

The likelihood function for the data can be expressed as

$$L = \prod_i \Phi(-[p^* - E(p_i|a_i, b_i)]/\sigma_{p|ab})^{\mathbb{1}(\text{outcome}_i=\text{positive})} \times \Phi([p^* - E(p_i|a_i, b_i)]/\sigma_{p|ab})^{\mathbb{1}(\text{outcome}_i=\text{negative})} \times \phi(a_i, b_i), \quad (16)$$

where $\sigma_{p|ab}$ is the standard deviation of p conditional on a and b,

$$\sigma_{p|ab} \equiv \sqrt{1 - \frac{1}{1-\rho_{ab}^2}\left[(\rho_{ap} - \rho_{bp}\rho_{ab})\rho_{ap} + (\rho_{bp} - \rho_{ap}\rho_{ab})\rho_{bp}\right]},$$

and the expectation of p_i conditional on a_i and b_i is

$$E(p_i|a_i, b_i) = \frac{1}{1-\rho_{ab}^2}\left[(\rho_{ap} - \rho_{bp}\rho_{ab})a_i + (\rho_{bp} - \rho_{ap}\rho_{ab})b_i\right].$$

We estimate the parameters ρ_{ap}, ρ_{ab}, ρ_{bp}, and p^* by maximizing the likelihood function in Equation (16).

As in the main text, I use a simulation study to examine the performance of the procedure. Table 4 presents these results. Not surprisingly, when the classifier that selected the test data is observed, the average areas under the inferred ROC curves are much closer to the average areas under the ROC curves that are based on random samples. We also see that the standard deviations of the areas under the inferred ROC curves are closer to the standard deviations of the areas under the ROC curves based on random samples.

References

Bradley, A. P. (1997): "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, 30(7), 1145–1159.

Breiman, L. (2001): "Random forests," *Machine Learning*, 45(1), 5–32.

Cortez, P., A. Cerdeira, F. Almeida, T. Matos, and J. Reis (2009): "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems*, 47(4), 547–553.

Crook, J., and J. Banasik (2004): "Does reject inference really improve the performance of application scoring models?," *Journal of Banking & Finance*, 28(4), 857–874.

Davis, J., and M. Goadrich (2006): "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240. ACM.

Dorfman, D. D., and E. Alf (1969): "Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals: Rating-method data," *Journal of Mathematical Psychology*, 6(3), 487–496.

Fawcett, T. (2006): "An introduction to ROC analysis," *Pattern Recognition Letters*, 27(8), 861–874.

Heckman, J. J. (1976): "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models," *Annals of Economic and Social Measurement*, 5(4), 475–492.

——— (1979): "Sample selection bias as a specification error," *Econometrica*, 47(1), 153–161.

Hernández-Orallo, J., P. Flach, and C. Ferri (2013): "ROC curves in cost space," *Machine Learning*, 93(1), 71–91.

Li, P., and M. Arshad Rahman (2011): "Bayesian analysis of multivariate sample selection models using gaussian copulas," *Advances in Econometrics*, 27, 269.

Liaw, A., and M. Wiener (2002): "Classification and Regression by randomForest," *R News*, 2(3), 18–22.

Moffat, A. (2013): "Seven numeric properties of effectiveness metrics," in *Information Retrieval Technology*, pp. 1–12. Springer.

Poirier, D. J. (1980): "Partial observability in bivariate probit models," *Journal of Econometrics*, 12(2), 209–217.

Van de Ven, W. P. M. M., and B. M. S. Van Praag (1981): "The demand for deductibles in private health insurance: A probit model with sample selection," *Journal of Econometrics*, 17(2), 229–252.

Wang, Y., L. Wang, Y. Li, D. He, and T.-Y. Liu (2013): "A Theoretical Analysis of NDCG Type Ranking Measures," in *Proceedings of the 26th Annual Conference on Learning Theory*.