

Aleksandra Baszczyńska *

EMPIRICAL AND KERNEL ESTIMATION OF THE ROC CURVE

Abstract. The paper presents chosen methods for estimating the *ROC* (*Receiver Operating Characteristic*) curve, including parametric and nonparametric procedures. Nonparametric approach may involve the use of empirical method or kernel method of the *ROC* curve estimation. In the analysis, an attempt to compare empirical and kernel *ROC* estimators is made, considering the impact of sample size, choice of smoothing parameter and kernel function in kernel estimation on the results of the estimation. Based on the results of simulation studies some suggestions, useful in the procedures of nonparametric *ROC* curve, are offered.

Keywords: *ROC* curve, empirical estimator, kernel method, smoothing parameter, kernel function.

1. INTRODUCTION

The *ROC* (*Receiver Operating Characteristic*) curve is a commonly used tool in economic analysis when different classification models are compared. Examples related to the economic phenomena are the following: enterprises division (threatened with collapse and non-threatened), workers division (threatened with dismissal and non-threatened), customers division (wishing to change the service provider and being loyal) or borrowers granted a consumer credit divisions (with defaulted loans and paid off loans). As a method of data visualization, the *ROC* curve comes from the technical diagnostics, especially in the electronic and signal theory, where its primary purpose has been connected with detection if signal can be treated as true or as noise. But a major research area of the *ROC* curve is the study in diagnostic medicine with assessing the accuracy of diagnostic tests in discriminating diseased from healthy patients. In these situations the *ROC* curve is an important decision support method.

Based on information about a set of objects, a division is made into one of two classes G_1 (objects with condition) and G_0 (objects without condition). *ROC* curve is used in the process of assessing the quality of the classification rules because this division may mean occurring the errors (the object is assigned

* Ph.D., Chair of Statistical Methods, University of Łódź.

to an incorrect class). The procedure of *ROC* curve allows to summarize distribution functions in two classes.

In assessing the value of prediction of decision rule, some measures are used. Let D be binary variable defining the presence of condition:

$$D = \begin{cases} 1 & \text{if condition is present,} \\ 0 & \text{if condition is absent.} \end{cases}$$

Let T be the result of the diagnostic test:

$$T = \begin{cases} 1 & \text{for positive test result,} \\ 0 & \text{for negative test result.} \end{cases}$$

The sensitivity of decision rule, defined as $SE = P(T = 1|D = 1)$, is the probability that the test result is positive, given that the condition is present. The specificity of the test $SP = P(T = 0|D = 0)$ is the probability that the test result is negative, given that the condition is absent. Sensitivity and specificity are used in the construction of *ROC* curve in such a way that the *ROC* curve is a plot of sensitivity associated with the test versus 1-specificity.

The *ROC* curve is defined as (cf. Fawcett (2006), Harańczyk (2010), Krzanowski, Hand (2009)):

$$ROC(p) = 1 - F_1(F_0^{-1}(1 - p)) \text{ for } 0 < p < 1 \quad (1)$$

where F_0 and F_1 are the distribution functions of class G_0 and G_1 respectively.

The distance between the *ROC* curve and the upper left corner $[0,1]$ is used in assessing the misclassification rate of the diagnostic test. The properties of the *ROC* curve are widely presented in Krzyśko *et al.* (2008).

2. ESTIMATION OF THE *ROC* CURVE

One-dimensional absolute continuous random variable X , called the diagnostic test variable, is used to assess if the object is classified to group G_0 and G_1 with the distribution function F_0 and F_1 respectively. In parametric approach we assume that the density function of the variable X is a mixture of two normal components. The parametric *ROC* curve estimator is the following:

$$\hat{R}_{Par}(p) = \Phi\left(\frac{\mu_1 - \mu_0}{\sigma_1} + \frac{\sigma_0}{\sigma_1} \Phi^{-1}(p)\right), \quad (2)$$

where μ_0, μ_1 are means, σ_0, σ_1 are standard deviations in mixture of normal distributions, $0 \leq p \leq 1$ and $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt$.

The simplest nonparametric estimator of the ROC curve is the empirical ROC curve estimator:

$$\hat{R}_{Emp}(p) = 1 - \hat{F}_{E1}(\hat{F}_{E0}^{-1}(1-p)), \quad (3)$$

where $\hat{F}_{Ej}(x) = \sum_{i=1}^{n_j} I_{(-\infty, x]}(X_{ji})$ for $j=0,1$ are empirical distribution function, $X_{0,1}, \dots, X_{0,n_0}$ and $X_{1,1}, \dots, X_{1,n_1}$ are independent samples from populations with F_0 and F_1 , respectively and $0 \leq p \leq 1$.

Empirical ROC curve estimator (3) is a step function on the unit square, and its jagged form is treated as its major drawback. Some trials to improve this estimator are presented in detail, for example, in Lloyd (2002) or Horová *et al.* (2012).

Another nonparametric ROC curve estimator is based on kernel method. This method was used for the first time in the procedure of density estimation, results in kernel density estimator $\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)$, where X_1, \dots, X_n is the random sample, $K(u)$ is kernel function and h is smoothing parameter.

A commonly used kernel function is Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$.

Choosing the smoothing parameter is a crucial problem in applying kernel methods, however there is no "optimal" value of this parameter in kernel procedures. Kernel functions and methods of choosing smoothing parameter are presented widely in literature. Kernel method is applied successfully in estimating distribution function, regression function and in testing hypotheses about independence or goodness-of-fit.

Kernel ROC curve estimator for $0 \leq p \leq 1$ has the form:

$$\hat{R}_{Ker}(p) = 1 - \hat{F}_{K1}(\hat{F}_{K0}^{-1}(1-p)), \quad (4)$$

where \hat{F}_{K_0} and \hat{F}_{K_1} are kernel estimators of distribution functions $\hat{F}_{K_j}(x, h_j) = \frac{1}{n_j} \sum_{i=1}^{n_j} W\left(\frac{x - X_{ji}}{h_j}\right)$ based on samples respectively $X_{0,1}, \dots, X_{0,n_0}$ and $X_{1,1}, \dots, X_{1,n_1}$ while $W(x) = \int_{-\infty}^x K(t) dt$ for kernel function $K(t) \geq 0$.

One of evaluating the classification model methods is the area under the *ROC* curve (denoted by *AUC*), defined as $AUC = \int_0^1 R(p) dp$. It takes the value from 0 to 1 while value close to 1 indicates high diagnostic accuracy (cf. Chrzanowski 2014; Domański, Pekasiewicz, Baszczyńska, Witaszczyk 2014; Misztal 2014).

3. RESULTS OF SIMULATION STUDY

The objective of the study is to compare chosen estimation methods of the *ROC* curve. An attempt is made to compare the performance of empirical and kernel *ROC* estimators, considering the impact of sample size, choice of smoothing parameter and kernel function in kernel estimation on the results of the estimation.

In the simulation study fifteen populations, introduced by Marron and Wand (1992), are taken into account. This collection of Gaussian mixture models, often used in works concerning the studies of performance of various kernels estimators (cf. Ruzgas, Drulyrè 2013), represents a wide range of density functions, including symmetric, asymmetric, unimodal and multimodal ones. Variety of distributions allows to regard different levels of similarity of populations taking into account.

From populations, samples of different sizes are drawn ($n = 10, 50, 100$). Estimators of *ROC* curve are calculated, treating samples from two specific populations from Marron and Wand's collection as group G_0 and G_1 . In the case of kernel estimator, Epanechnikov kernel function and method of maximal smoothing parameter are used (cf. Horová *et al.* 2012). In this way the estimators of the *ROC* curve are used in the process of distinguishing two populations. When the density functions of the populations are similar, the distance between the estimator and the diagonal line should be small; otherwise this distance should be bigger.

The chosen results of the first stage of empirical study when the sample size is taken into consideration are presented in Figure 1. G_0 is a sample from the population with a normal distribution and G_1 is a sample from the symmetric but multimodal population (10. population in Marron and Wand collection).

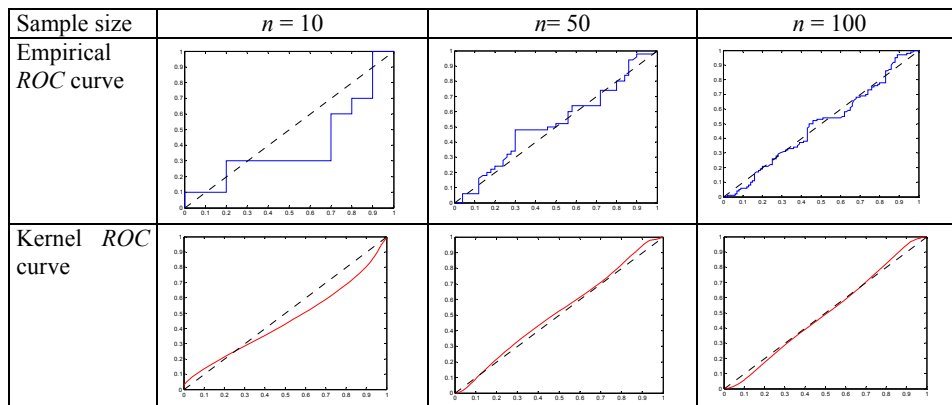


Figure 1. Empirical and kernel *ROC* curve estimators for different samples sizes when G_0 is a sample from normal distribution and G_1 is a sample from population 10 (symmetric, multimodal)
Source: own elaboration.

It can be seen that in the case of both empirical *ROC* curve estimator and kernel one, the closeness of estimators and the diagonal line is small, what can indicate that test is not usable for separation of regarded objects (populations are similar), though the difference between density functions is easy to notice and in fact they are two different populations. The bigger the sample size is, the smaller the closeness to diagonal line is.

In the second stage of study, two populations from the collection are specified, for which the estimators are calculated. The chosen results (estimators and *AUC* values) for $n = 50$ are presented.

When the asymmetry becomes stronger in populations from which samples G_1 are generated (for example $l = 2$ and $l = 3$) the closeness between *ROC* curve estimators and the upper left corner becomes bigger. It can mean that asymmetry is such a characteristic of random variable which causes that it is easy to detect the difference between populations using empirical or kernel estimators. When modality is the main characteristic that differs two populations (for example $l = 6, \dots, 15$) the *ROC* curve estimators should not be used in detecting the differences between populations. It can be noticed that the bigger the number of modes is, the smaller the closeness to diagonal line is.

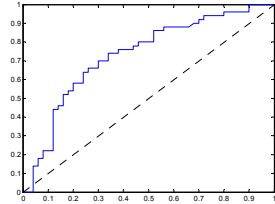
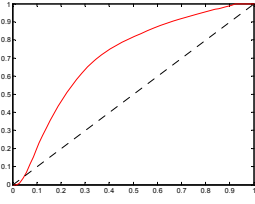
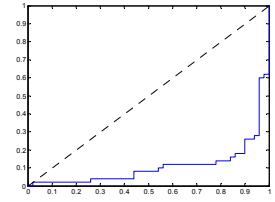
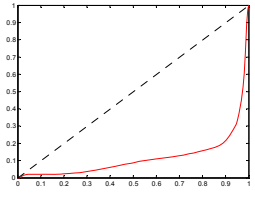
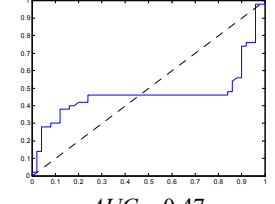
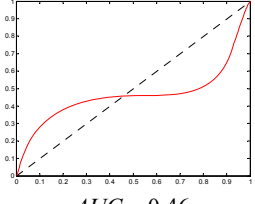
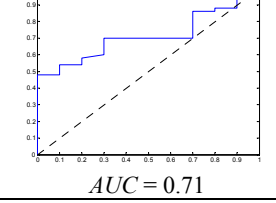
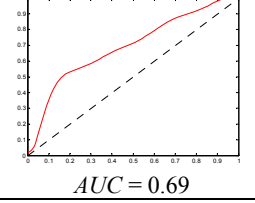
Distributions of populations from which the samples are generated	Empirical ROC estimator	Kernel ROC estimator
$N(0,1)$ and ($l=2$) $\frac{1}{5}N(0,1) + \frac{1}{5}N\left(\frac{1}{2}, \frac{2}{3}\right) + \frac{3}{5}N\left(\frac{13}{12}, \frac{5}{9}\right)$	 $AUC = 0.73$	 $AUC = 0.71$
$N(0,1)$ and ($l=3$) $\sum_{i=0}^7 \frac{1}{8} N\left(3\left[\left(\frac{2}{3}\right)^i - 1\right], \left(\frac{2}{3}\right)^i\right)$	 $AUC = 0.11$	 $AUC = 0.10$
$N(0,1)$ and ($l=7$) $\frac{1}{2}N\left(-\frac{3}{2}, \frac{1}{2}\right) + \frac{1}{2}N\left(\frac{3}{2}, \frac{1}{2}\right)$	 $AUC = 0.47$	 $AUC = 0.46$
$N(0,1)$ and ($l=8$) $\frac{3}{4}N(0,1) + \frac{1}{4}N\left(\frac{3}{2}, \frac{1}{3}\right)$	 $AUC = 0.71$	 $AUC = 0.69$

Figure 2. Empirical and kernel ROC curve estimators and AUC values for G_0 – sample from normal distribution and G_1 – sample from population l ($l = 2, 3, 7, 8$), $n = 50$

Source: own elaboration.

4. CONCLUSIONS

Based on the results of simulation studies it can be stated that both empirical and kernel ROC curve estimators behave in similar way. The area under the estimator (AUC) is, in most cases, smaller for kernel estimator, what is closely connected with jagged form of empirical estimator. The results indicate that kernel estimator may be treated as more cautious procedure what is, in fact, an

advantage for novice users of statistical procedures. Estimators of ROC curve are recommended especially in situations when strong population asymmetry is suspected. Regarded nonparametric procedures for estimation the ROC curve are easy to implement because of special computation programs and should be used instead of parametric approaches. They do not assume the density function, so can be useful when the researcher has no additional information about population. In further researches the emphasis should be placed on comparing kernel estimators with different values of smoothing parameter and kernel functions.

REFERENCES

- Chrzanowski M. (2014), Weighted Empirical Likelihood Inference for the Area under the ROC Curve, *Journal of Statistical Planning and Inference*, 147, 159–172.
- Domański C., Pekasiewicz D., Baszczyńska A., Witaszczyk A. (2014), *Testy statystyczne w procesie podejmowania decyzji*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.
- Fawcett T. (2006), An Introduction to ROC Analysis, *Pattern Recognition Letters*, 27, 861–874.
- Harańczyk G. (2010), *Krzywe ROC, czyli ocena jakości klasyfikatora i poszukiwanie optymalnego punktu odcięcia*, Statsoft Polska, www.statsoft.pl/czytelnia.html.
- Horová I., Koláček J., Zelinka J. (2012), *Kernel Smoothing in Matlab. Theory and Practice of Kernel Smoothing*, World Scientific, New Jersey.
- Krzanowski W., Hand D. (2009), *ROC Curves for Continuous Data*, CRC Press.
- Krzyśko M., Wołyński W., Górecki T., Skorzybut M. (2008), *Systemy uczące się. Rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości*, Wydawnictwa Naukowo-Techniczne, Warszawa.
- Lloyd C. (2002), Estimation of a Convex ROC Curves, *Statistics and Probability Letters*, 59, 1, 99–111.
- Marron J., Wand M. (1992), Exact Mean Integrated Squared Error, *The Annals of Statistics*, 20, 2, 712–736.
- Misztal M. (2014), On the Selected Methods for Evaluating Classification Models, *Acta Universitatis Lodzianis Folia Oeconomica*, 3 (302), 161–173.
- Ruzgas T., Drulyrė I. (2013), Kernel Density Estimation for Gaussian Mixture Models, *Lithuanian Journal of Statistics*, 52, 1, 14–21.

Aleksandra Baszczyńska

EMPIRYCZNY I JĄDROWY ESTYMATOR KRZYWEJ ROC

Streszczenie. W pracy rozważane są wybrane metody estymacji krzywej ROC (*Receiver Operating Characteristic*), w tym metody parametryczne i nieparametryczne. Podejście nieparametryczne może oznaczać zastosowanie empirycznego estymatora krzywej ROC lub estymatora jądrowego. Podjęta jest próba porównania estymacji empirycznej oraz jądrowej ze szczególnym uwzględnieniem wpływu liczebności próby, jak również metody wyboru parametru wygładzania i funkcji jądra na rezultat procedury estymacyjnej. W oparciu o wyniki badania symulacyjnego określone są wskazówki użyteczne w procedurach estymacji krzywej ROC.

Słowa kluczowe: krzywa ROC, funkcja jądra, parametr wygładzania.

