POLITECNICO DI TORINO
Repository ISTITUZIONALE

On the definition of a concentration function relevant to the ROC curve

Original
On the definition of a concentration function relevant to the ROC curve / Gasparini, M., Sacchetto, L.. - In: METRON. - ISSN 0026-1424. - 78:3(2020), pp. 271-277. [10.1007/s40300-020-00191-5]

Availability:
This version is available at: 11583/2854373 since: 2020-12-01T21:20:28Z

Publisher:
Springer-Verlag Italia s.r.l.

Published
DOI:10.1007/s40300-020-00191-5

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

30 October 2026

METRON (2020) 78:271–277
https://doi.org/10.1007/s40300-020-00191-5

On the definition of a concentration function relevant to the ROC curve

Mauro Gasparini[1] · **Lidia Sacchetto**[2]

Received: 31 December 2019 / Accepted: 7 October 2020 / Published online: 23 October 2020
© The Author(s) 2020

Abstract
This work provides a definition of concentration curve alternative to the one presented on this journal by Schechtman and Schechtman (Metron 77:171–178, 2019). Our definition clarifies, at the population level, the relationship between concentration and the omnipresent ROC curve in diagnostic and classification problems.

Keywords Likelihood ratio · Lorenz curve · Length-Biased · Gini

1 A critical appraisal of a paper by E. Schechtman and G. Schechtman

In a paper appeared recently on this journal Schechtman and Schechtman [6] try to shed some light on the relationship between the Gini Mean Difference (Gini), the Gini Covariance (co-Gini), the Lorenz curve, the Receiver Operating Characteristic (ROC) curve and a particular definition of concentration function. The purpose of the paper is commendable, since there is a lot of confusion regarding the various relationships among these concepts. In particular, we agree that the ROC curve and its functions (such as the Area Under the Curve, AUC), as well as an appropriate definition of relative concentration of a probability distribution with respect to another, are bivariate objects tying together two different distributions, and can not be reduced to univariate indices such as the Gini.

Schechtman and Schechtman [6] build on the wealth of research reviewed in the monograph by Yitzhaki and Schechtman [7], where a whole technology based on the Gini and the co-Gini are proposed as basic tools to study variability, correlation, regression and the like. In particular, the authors try to use certain conditional expectations to establish the connection between concentration and ROC. In this note, we claim their approach is not justified in the

✉ Mauro Gasparini
mauro.gasparini@polito.it

Lidia Sacchetto
lidia.sacchetto@polito.it

[1] Department of Mathematical Sciences “G.L. Lagrange”, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10124 Torino, Italy

[2] Department of Mathematical Sciences “G.L. Lagrange”, Politecnico di Torino and Università di Torino, Corso Duca degli Abruzzi 24, 10124 Torino, Italy

diagnostic (classification) setup, where ROC curves typically arise, and propose an alternative simpler connection between concentration and ROC curves based on first principles, namely the likelihood ratio and the application of the Neyman-Pearson lemma.

Studying how jointly distributed random variables interrelate is a very fundamental problem in Statistics and its applications to Economics and the Sciences. However, when turning to the diagnostic (or classification) setup, one typically observes one or more diagnostic variables (called features in the Machine Learning literature) from two populations and try to set up a rule that discriminates between them. Some special requirements can then be identified:

1. Two probability distributions should be evaluated as distinct explanations of the data, rather than from a joint point of view; for example, a diagnostic marker Y observed on a sick patient will have a different distribution from the same diagnostic marker X observed on a healthy patient, and in no way the same marker can be observed jointly under both the sick and the healthy conditions. Conditioning on the population label is possible (as done, for example, in the causal inference literature), but not conditioning of, say, Y on X.
2. The definition of the ROC curve and the associated concentration function should be viable also in the multivariate setup; for example, more than one diagnostic marker can be observed on the same patient.
3. The definition of the ROC curve and the associated concentration function must be given both at the population and at the sample level, as widely discussed in the ROC literature (see for example [3]); a clear definition of the ROC curve at the population level is necessary to understand basic ideas and to give appropriate definitions.

We claim the definition of concentration curve contained in [6] is not appropriate for the diagnostic setup since:

a. Conditional distributions are used in the Definition 1 of [6], thus contradicting requirement (1);
b. Percentiles are used in the same definition, thus contradicting requirement (2);
c. In [6], the discussion on the ROC curve is mantained at the sample level only, making it hard to understand what is, for example, the definition of population ROC curve.

2 The ROC curve of the likelihood ratio test, a definition of concentration curve and their equivalence

We now briefly introduce the diagnostic (i.e. classification) setup. Assume that Y is a continuous random variable with distribution function F_Y and strictly positive density f_Y and X is a continuous random variable with distribution function F_X and strictly positive density f_X. As mentioned in the previous section, let Y and X represent, respectively, the relevant diagnostic variable under the two conditions to be compared by a diagnostic test. For example, Y may be a biological marker measured in a diseased person, whereas X is the same marker when measured in a healthy person.

Now suppose a new observation Z, coming in un unknown way from one of the two populations, has to be assigned either to the X or to the Y population based on some function of $s(Z)$, called the score. Usually, the score is real valued and the decision rule is worded as "assign Z to the Y population if $s(Z)$ is larger than a threshold t". The ROC curve of the decision rule is then the locus of the points obtained by varying the threshold t:

$$\{\mathrm{FPR}(t), \mathrm{TPR}(t), t \in \mathcal{R}\},$$

where the false positive rate FPR is the probability the decision rule assigns the object to the Y population given the object comes from the X population and the true positive rate TPR is the probability the decision rule assigns the object to the Y population given the object comes from the Y population.

If, for the sake of simplicity, we assume F_X and F_Y are known, it is well known that the best decision rule is the one using as score the likelihood ratio itself, as discussed for example in [8]. Optimality stems from the Neyman-Pearson lemma—and from its Dantzig-Wald generalization ([2]), as it was recognized very early in the ROC literature. Such a decision rule is

Definition 1 The likelihood ratio based test assigns Z to the Y (resp. X) population if $f_Y(Z)/f_X(Z) > t$ (resp. $\leq$) for some t.

If F_X and F_Y are not known, then they must be estimated based on two samples from the X and Y populations; this problem, often called supervised learning classification, populates a large amount of Statistics and Machine Learning literature, but it is beyond the scope of this article.

To obtain an explicit formula for the ROC curve of the optimal likelihood ratio based test, define the likelihood ratio random variables

$$L_X = \frac{f_Y(X)}{f_X(X)} \quad \text{and} \quad L_Y = \frac{f_Y(Y)}{f_X(Y)}$$

which are pro bono random variables since they are functions of X and Y, respectively. Now, for the sake of simplicity, assume L_X is continuous with distribution function $H_X(l) = \mathrm{P}_X(L_X \leq l) = \mathrm{P}_X(L_X < l), l > 0$ which has inverse $H_X^{-1}(\cdot)$, its quantile function. Similarly, assume L_Y is a continuous random variable with distribution function $H_Y(\cdot)$. Then it is easy to see that

$$\mathrm{TPR}_{LR}(t) = \mathrm{P}_Y\left(\frac{f_Y(Y)}{f_X(Y)} > t\right) = 1 - H_Y(t)$$
$$\mathrm{FPR}_{LR}(t) = \mathrm{P}_X\left(\frac{f_Y(X)}{f_X(X)} > t\right) = 1 - H_X(t)$$

so that, eliminating t and setting $q = \mathrm{FPR}_{LR}$, the ROC curve of the likelihood ratio based diagnostic test can be written in explicit form:

$$\mathrm{ROC}_{LR}(q) = 1 - H_Y(H_X^{-1}(q)) \qquad 0 \leq q \leq 1, \tag{1}$$

while, by definition, $\mathrm{ROC}_{LR}(0) = 0$ and $\mathrm{ROC}_{LR}(1) = 1$.

We now turn to an appropriate definition of concentration curve in this situation.

Definition 2 The concentration curve of Y with respect to X is the function $\varphi(p)$, $p \in [0, 1]$ such that $\varphi(0) = 0$, $\varphi(1) = 1$ and

$$\varphi(p) = H_Y(H_X^{-1}(p)), \quad p \in (0, 1). \tag{2}$$

Such definition is a special case of [1] (for details, see [5]); according to their suggestion, for each $p \in [0, 1]$ the concentration function $\varphi(p)$ is the likelihood ratio Y-mass of a set collecting the smallest p fraction of the likelihood ratio X-mass.

The main point of this work is to notice the obvious relationship between the two definitions given above; the likelihood ratio based test has a ROC curve which is a bijective transformation of the concentration function in Definition 2:

$$\mathrm{ROC}_{LR}(q) = 1 - \varphi(1 - q) \qquad 0 \leq q \leq 1, \tag{3}$$

Notice that the ROC of the likelihood ratio based test is proper, i.e. a nondecreasing, continuous and concave function, while other ROC curves not based on the likelihood ratio may not be. Similarly, the concentration curve based on the likelihood ratio is nondecreasing, continuous and convex (detailed proofs can be found in [1]), as concentrations are usually required to be. The likelihood ratio is a necessary requirement for these constructions, and we claim likelihood ratios, and not conditional expectations as in [6], are the proper tools to establish the connection between the two objects.

3 The Lorenz curve and the AUC of the likelihood ratio based test

An interesting special case discussed in [1] arises when X is a positive random variable with finite mean $\mu_x = \int_0^\infty x f_X(x)dx$ and Y is the length-biased version of X, i.e.

$$f_Y(y) = \frac{y f_X(y)}{\mu_X}, \quad y > 0.$$

In economic applications, Y represents wealth; in general, it may be a transferable character, i.e. some characteristic which can in theory be transported from one unit of the population to another. This is the Lorenz-Gini setup. The likelihood ratios simplify to

$$L_X = \frac{f_Y(X)}{f_X(X)} = \frac{X f_X(X)}{\mu_X f_X(X)} = \frac{X}{\mu_X}$$

and

$$L_Y = \frac{f_Y(Y)}{f_X(Y)} = \frac{Y f_X(Y)}{\mu_X f_X(Y)} = \frac{Y}{\mu_X}$$

so that $H_X(l) = F_X(\mu_x l)$ and $H_Y(l) = F_Y(\mu_x l)$ and finally

$$\varphi_{Lorenz}(p) = H_Y(H_X^{-1}(p)) = F_Y(F_x^{-1}(p)) = \frac{\int_0^{F_x^{-1}(p)} y f_X(y)dy}{\int_0^\infty x f_X(x)dx},$$

in which we recognize one of the usual forms of the Lorenz curve. We have just proven the following

Lemma 1 *In the Lorenz-Gini scenario, i.e. when $f_Y(y) = y f_X(y)/\mu_X$, the concentration curve is the usual Lorenz curve.*

A second important consequence of our definitions in the previous section is about the AUC of the likelihood ratio based test, which can be easily computed as follows:

$$\text{AUC}_{LR} = \int_0^1 ROC_{LR}(q)dq = \int_0^1 (1 - \varphi(1-q))dq = 1 - \int_0^1 \varphi(s)ds. \tag{4}$$

Now, in the Lorenz-Gini scenario, the Gini concentration coefficient (Gini) is defined to be twice the area between the diagonal and the Lorenz curve:

$$\text{Gini} = 2\int_0^1 (p - \varphi_{Lorenz}(p))dp = 1 - 2\int_0^1 \varphi_{Lorenz}(p)dp$$

Since the concentration curve is a generalization of the Lorenz curve which describes the concentration of one variable with respect to another (and not necessarily its length-biased

version), we can define the generalized Gini as

$$\text{Gini}_{gen} = 2\int_0^1 (p - \varphi(p))dp,$$

similarly to the co-Gini in [6]. Substituting into expression (4) we obtain the following corollary.

Corollary 1 *The AUC of the optimal likelihood ratio based diagnostic test equals*

$$AUC_{LR} = \frac{1}{2}\big(1 + \text{Gini}_{gen}\big).$$

The same result can be found in [4] and mentioned by several other authors. We stress that the result is true for the likelihood ratio based test and, of course, for models with monotone likelihood ratios (like the example considered in [4]) but not in general for the AUC of any ROC.

4 Two examples

Example 1 Let X be exponential with rate parameter λ_X and Y be exponential with rate parameter λ_Y and assume, as it is customary, that $\lambda_X > \lambda_Y$, so that Y is stochastically greater than X (this corresponds to a situation where the greater a diagnostic marker, the more is indicative of disease). Then it is easy to verify that

$$H_X(l) = \text{P}\left(\frac{f_Y(X)}{f_X(X)} \le l\right) = \text{P}\left(\frac{\lambda_Y e^{-\lambda_Y X}}{\lambda_X e^{-\lambda_X X}} \le l\right) = 1 - \left(\frac{1}{rl}\right)^{r/(r-1)}$$

for $l > 1/r$ and 0 otherwise, where $r = \lambda_X/\lambda_Y$. Similarly,

$$H_Y(l) = \text{P}\left(\frac{f_Y(Y)}{f_X(Y)} \le l\right) = 1 - \left(\frac{1}{rl}\right)^{1/(r-1)}$$

for $l > 1/r$ and 0 otherwise. Also,

$$H_X^{-1}(p) = \frac{1}{r}\left(\frac{1}{1-p}\right)^{(r-1)/r}$$

so that the concentration function is

$$\varphi(p) = H_Y(H_X^{-1}(p)) = 1 - (1-p)^{1/r}, \quad p \in (0,1),$$

the ROC curve of the likelihood ratio based optimal test is

$$\text{ROC}_{LR}(q) = q^{1/r} \qquad 0 \le q \le 1,$$

and

$$\text{AUC}_{LR} = \frac{r}{r+1}.$$

Example 2 Let X be exponential with rate parameter λ_X and assume Y is its length-biased version, so that

$$f_Y(y) = \frac{y\lambda_X e^{-\lambda_X y}}{1/\lambda_X} = \lambda_X^2 y e^{-\lambda_X y}, \quad y > 0,$$

i.e. Y is a gamma random variable with parameters 2 and λ_X. This is a Lorenz-Gini scenario, where it is easy to verify that

$$H_X(l) = \mathrm{P}\left(\frac{f_Y(X)}{f_X(X)} \leq l\right) = \mathrm{P}\left(\lambda_X X \leq l\right) = 1 - e^{-l}$$

whereas, after some calculus,

$$H_Y(l) = \mathrm{P}(\frac{f_Y(Y)}{f_X(Y)} \leq l) = \mathrm{P}(\lambda_X Y \leq l) = 1 - e^{-l} - le^{-l}. \tag{5}$$

Since $H_X^{-1}(p) = -\log(1-p)$,

$$\varphi(p) = p + (1-p)\log(1-p), \quad \mathrm{ROC}_{LR}(q) = q - q\log(q).$$

5 Conclusions

The definition of concentration function given here is a convenient one since it compares two alternative probability distributions, reaching a natural bivariate generalization of the Lorenz curve. The discussion on the concentration and the ROC curves at the population level allows for a deeper understanding of the concepts, including the centrality of the likelihood ratio. In higher dimensions, computations may become very hard, but all results apply nonetheless. In particular, the likelihood ratio may then be an efficient dimension reduction technique which reduces the comparison to a one-dimensional problem and allows for Definition 2 of concentration function without involving higher dimensional conditional expectations or quantiles.

We hope we have convinced the reader that the nature of the diagnostic (classification) problem requires a definition of concentration function which does not involve conditional and joint distributions of the populations which are being compared.

Acknowledgements This study was funded by the Italian Ministry of Education, University and Research, MIUR, grant Dipartimenti di Eccellenza 2018–2022 (E11G18000350001).

Funding Open access funding provided by Politecnico di Torino within the CRUI-CARE Agreement.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Cifarelli, D.M., Regazzini, E.: On a general definition of concentration function. SANKHYA B **49**, 307–319 (1987)

2. Dantzig, G.B., Wald, A.: On the fundamental lemma of Neyman and Pearson. Ann. Math. Stat. **22**(1), 87–93 (1951)
3. Krzanowski, W.J., Hand, D.J.: ROC Curves for Continuous Data. Chapman & Hall, London (2009)
4. Lee, W.C.: Probabilistic analysis of global performances of diagnostic tests: interpreting the Lorenz curve-based summary measures. Stat. Med. **18**, 455–471 (1999)
5. Sacchetto, L. and Gasparini, M.: Proper likelihood ratio based ROC curves for general binary classification problems. arXiv:1809.00694 (2018)
6. Schechtman, E., Schechtman, G.: The relationship between Gini terminology and the ROC curve. Metron **77**, 171–178 (2019)
7. Yitzhaki, S., Schechtman, E.: The Gini Methodology. Springer, Berlin (2012)
8. Zou, K.H., Liu, A., Bandos, A.I., Ohno-Machado, L., Rockette, H.E.: Statistical Evaluation of Diagnostic Performance Topics in ROC Analysis. Chapman & Hall, London (2012)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.