Economics Bulletin

Volume 39, Issue 2

A note on Gini Principal Component Analysis

Téa Ouraga
Université de Nîmes - Laboratoire CHROME

Abstract

In this paper, a principal component analysis based on the Gini index - Gini PCA - is proposed in order to deal with contaminated samples. The operator underlying the Gini index is a covariance-based operator, which provides a l1 metric well suited for dealing with outliers. It is shown, with simple Monte Carlo experiments, that the results of the standard Principal Component Analysis (PCA) may be drastically affected whereas some robustness holds with Gini PCA.

This paper has been presented at the GLADAG 2017 conference in Milan. I would like to thank the organizer, Pr Greselin, my director of thesis Pr S. Mussard and the participants for helpful comments. Remaining errors are mine.
Citation: Téa Ouraga, (2019) "A note on Gini Principal Component Analysis", *Economics Bulletin*, Volume 39, Issue 2, pages 1076-1083
Contact: Téa Ouraga - jeromeouraga@gmail.com.
Submitted: March 07, 2019. **Published:** May 02, 2019.

1 - Introduction

In 1912, Gini proposed the Gini Mean Difference index (GMD) as a new way to measure inequality and disparity between individuals in a given population:

$$GMD_x = \mathbb{E}\left|x_i - x_j\right|, \tag{1}$$

where x_i and x_j are two realizations of the random variable x. The GMD is based on the taxi-cab distance and thus offers an alternative measure of the usual variance based on the euclidean metrics:

$$\sigma_x^2 = \mathrm{cov}(x,x) = \frac{1}{2}\,\mathbb{E}\left|x_i - x_j\right|^2. \tag{2}$$

Since then, two main approaches have been developed in the literature for analyzing the variability between two random variables.

The first one is based on the covariance between the c.d.f. of the random variable x and that of y, expressed as:

$$S_{x,y} = \mathrm{cov}(F(x), F(y)). \tag{3}$$

This is the well-known Spearman's method defined to be the rank method. The second one, the Gini approach, has been developed by Schechtman and Yitzhaki (2003) who paved the way on the covariance Gini operator – cogini operator from now on – which can be seen as a mixture of the variance and Spearman's pure rank approaches:

$$cog(x,y) = \mathrm{cov}(x, F(y)) \;;\; cog(y,x) = \mathrm{cov}(y, F(x)). \tag{4}$$

It is noteworthy that $cog(x,x) = 1/4GMD_x$, as a consequence, the cogini operator is closely related to the ℓ_1 metric.

The cogini operator has some appealing features. For instance, Olkin and Yitzhaki (1992) and Yitzhaki and Schechtman (2013) point out that the ordinary least squares method can be employed by replacing the usual covariance operator by the cogini one. Their Gini regression has been shown to be robust to outliers. Indeed, the variance criterion may be misleading to handle a sample with extreme values or to deal with heavy-tailed distributions, see Carcea and Serfling (2015) in the case of times series. Also, as shown by Greselin (2015) the use of cogini operators close to Choquet integrals may be useful to unify measures of inequality and risk.

In this paper, we start from the recognition that the cogini lies in the family of robust statistics, and as such, it is a good candidate to perform Principal Component Analysis in the Gini sense (Gini PCA). In the field of PCA, Baccini *et al.* (1996) were among the first authors dealing with a ℓ_1-norm PCA framework. Their idea was to robustify the standard PCA by means of the Gini Mean Difference as an estimator of the standard deviation. Ding *et al.* (2006) made use of the R_1 norm to robustify the PCA, in which the Euclidean distance is applied over the dimensions of the matrix only, whereas Frobienus norm is concerned with the Euclidean distance applied to both dimensions and observations (rows of the matrix of the data).

The aim of this paper is to use the cogini operator underlying the correlation Gini index in order to provide a Gini PCA less sensitive to outlying observations than the usual PCA by substituting the variance-covariance matrix to the Gini correlation matrix. Contrary to Baccini *et al.* (1996) in which the PCA is formalized by replacing the standard deviation of each variable by its GMD, we employ the Gini correlation index between all pairs of variables (Section 2). We show with simple Monte Carlo simulations that the Gini PCA is robust to outliers thanks to the relative and absolute contributions, that are respectively, the distance of the observations to the principal components and their contributions to the overall Gini correlation (Section 3). Section 4 closes the note.

2 - Gini PCA

Let $\mathbf{X} \equiv [x_{ik}]$ be a $N \times K$ matrix that describes N observations on K dimensions such that $N \gg K > 1$, with elements $x_{ik} \in \mathbb{R}$ that reports the score of observation i on dimension k. The $N \times 1$ vectors representing each column of $\mathbf{X}$ are expressed as $\mathbf{x}_k$, for all $k \in \{1, \ldots, K\}$, such that $\mathbf{x}_k \neq c\mathbf{1}_K$, with c a real constant (and $\mathbf{1}_K$ a column vector of ones of dimension K). The ℓ_1 norm of $\mathbf{x}_k$ is given by $\|\mathbf{x}_k\|_1 = \sum_{i=1}^{N} |x_{ik}|$. The arithmetic means of the variables are given by $\bar{\mathbf{x}}_k$. An estimator of the Gini Mean Difference between two variables $\mathbf{x}_\ell$ and $\mathbf{x}_k$, proposed by Schechtman and Yitzhaky (1987), is given by:

$$GMD(\mathbf{x}_\ell, \mathbf{x}_k) := \frac{4}{N} \sum_{i=1}^{N} (x_{i\ell} - \bar{\mathbf{x}}_\ell)(\hat{F}(x_{ik}) - \bar{F}_{\mathbf{x}_k}), \tag{5}$$

where $\hat{F}(x_{ik})$ is the estimated cumulative distribution function of $\mathbf{x}_k$ at point i, $\bar{F}_{\mathbf{x}_k}$ its mean, with $\ell, k = 1, \ldots, K$. When $k = \ell$ the GMD represents the variability of the variable $\mathbf{x}_\ell$ with itself, see Eq.(1). Alternatively, it is possible to define the rank vector $R(\mathbf{x}_\ell)$ of variable $\mathbf{x}_\ell$ as an estimator of $F(\mathbf{x}_\ell)$,

$$\hat{F}(x_{i\ell}) = \frac{R(x_{i\ell})}{N} := \begin{cases} \frac{\#\{x \leq x_{i\ell}\}}{N} & \text{if no ties} \\ \frac{\sum_{i=1}^{p} \#\{x \leq x_{i\ell}\}}{Np} & \text{if } p \text{ ties } x_{i\ell}. \end{cases} \tag{6}$$

The rank vector assigns the value 1 to the smallest value of vector $\mathbf{x}_\ell$, and so on. In the case of ties, the mean rank is computed as shown below for the first observation:

$$\mathbf{x}_\ell = \begin{pmatrix} 1 \\ 1 \\ 4 \\ 7 \\ 6 \end{pmatrix} \longrightarrow R(\mathbf{x}_\ell) = \begin{pmatrix} 1,5 \\ 1,5 \\ 3 \\ 5 \\ 4 \end{pmatrix} \tag{7}$$

A bias corrected estimator of GMD is,

$$GMD(\mathbf{x}_\ell, \mathbf{x}_k) := \frac{4}{N(N-1)} \sum_{i=1}^{N} (x_{i\ell} - \bar{\mathbf{x}}_\ell)(R(x_{ik}) - \bar{R}_{\mathbf{x}_k}), \ \forall k, \ell = 1, \ldots, K, \tag{8}$$

with $\bar{R}_{\mathbf{x}_k}$ the mean of the rank vector of variable $\mathbf{x}_k$. The Gini correlation coefficient, the G-correlation frown now on, is defined as follows:

$$GC(\mathbf{x}_\ell, \mathbf{x}_k) := \frac{GMD(\mathbf{x}_\ell, \mathbf{x}_k)}{GMD(\mathbf{x}_\ell, \mathbf{x}_\ell)}, \tag{9}$$

with $GC(\mathbf{x}_k, \mathbf{x}_k) = 1$ for all $k = 1, \ldots, K$. Following Yitzhaki (2003), the G-correlation is well-suited for the measurement of correlations in the case of distributions with atypical points and in general in the case of non-normal distributions.

Property – Yitzhaki (2003):
(i) $-1 \leq GC(\mathbf{x}_\ell, \mathbf{x}_k) \leq 1$.
(ii) *If the variables* $\mathbf{x}_\ell$ *and* $\mathbf{x}_k$ *are independent, for all* $k \neq \ell$, *then* $GC(\mathbf{x}_\ell, \mathbf{x}_k) = GC(\mathbf{x}_k, \mathbf{x}_\ell) = 0$.
(iii) *For any given monotonic transformation* φ, $GC(\mathbf{x}_\ell, \varphi(\mathbf{x}_k)) = GC(\mathbf{x}_\ell, \mathbf{x}_k)$ [*in the same than Spearman's coefficient*].
(iv) *For any given linear transformation* φ, $GC(\varphi(\mathbf{x}_\ell), \mathbf{x}_k) = GC(\mathbf{x}_\ell, \mathbf{x}_k)$ [*in the same manner than Pearson's coefficient*].
(v) *If* $\mathbf{x}_k$ *and* $\mathbf{x}_\ell$ *are exchangeable up to a linear transformation, then* $GC(\mathbf{x}_\ell, \mathbf{x}_k) = GC(\mathbf{x}_k, \mathbf{x}_\ell)$.

Another property could have been added to the previous ones, that of robustness of the G-correlation index. In order to assess its robustness in PCA frameworks, let us define the $K \times K$ G-correlation matrix containing the Gini correlations between each and every pairs of variables:

$$GC(\mathbf{X}) \equiv [GC(\mathbf{x}_\ell, \mathbf{x}_k)]. \tag{10}$$

Let $\mathbf{X}^c$ and $\mathbf{R}^c$ be the $N \times K$ column matrices containing respectively, the centered $\mathbf{x}_k$ vectors and the centered rank vectors. On the other hand, let the matrix of basis vectors be $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_h]$ containing the the eigen vector in columns such that there exist h non-zero eigen values $\lambda_1, \ldots, \lambda_h$.

Proposition *The Gini PCA consists in solving for the eigen values* λ_k *for all* $k = 1, \ldots, K$ *that maximize the Gini variability of* $\mathbf{X}$ *such that:*

$$\lambda_k = \arg\max \mathbf{b}_k^\intercal GC(\mathbf{X}) \mathbf{b}_k = \arg\max \mathbf{b}_k^\intercal (\mathbf{X}^c)^\intercal \mathbf{R}^c \mathbf{b}_k. \tag{11}$$

The eigen values λ_k are derived from the Gini correlation matrix instead of the usual variance-covariance matrix. The eigen vectors are normalized such that $\|\mathbf{b}_k\|_1 = 1$. Then, the observations are projected such that $\mathbf{F} = \mathbf{X}^c \mathbf{B}$, with $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_h]$ the matrix of projected observations into the new subspace spanned by the eigen vectors $\mathbf{b}_k$.

Baccini *and al.* (1996) proposed a ℓ_1 PCA solely based on the diagonal terms of $GC(\mathbf{X})$. In our approach, the extra-diagonal terms, representing the Gini correlation between the variables $\mathbf{x}_k$ and $\mathbf{x}_\ell$ ($k \neq \ell$), are taken into account in order to attenuate the influence of the outliers that could occur in those correlations as well.

3 - Monte Carlo Simulations

As in the standard PCA, the results of the Gini PCA may be interpreted thanks to absolute contributions (ACT) and relative contributions (RCT). ACT_{ik} represents the share of axis $\mathbf{f}_k$ variability (in the Gini sense) captured by each observation i. This statistics the number of significant components (axis) $\mathbf{f}_k$ to be selected. RCT_{ik} is the distance of observation i towards a component $\mathbf{f}_k$.

Definition 3.1 *The absolute contribution of an individual i to the Gini variability of a principal component $\mathbf{f}_k$ is:*

$$ACT_{ik} = \frac{f_{ik} r_{ik}}{GMD(\mathbf{f}_k, \mathbf{f}_k)}, \ \forall k = 1, \ldots, h, \tag{12}$$

where r_{ik} is the rank of individual i on the principal axis $\mathbf{f}_k$ and f_{ik} the score of observation i on component $\mathbf{f}_k$.

The absolute contribution of each i to the Gini mean difference of $\mathbf{f}_k$ is such that $ACT_{ik} \in [0, 1]$ and $\sum_{i=1}^{N} ACT_{ik} = 1$.

Definition 3.2 *The relative contribution of an individual i to a component $\mathbf{f}_k$ is:*

$$RCT_{ik} = \frac{|f_{ik}|}{\|\mathbf{f}_i\|_1}, \ \forall k = 1, \ldots, h, \tag{13}$$

where $\mathbf{f}_i$ is the i-th row of matrix $\mathbf{F}$.

On the one hand, Monte Carlo experiments are conducted with 5-variate normal distributions of size $N = 500$ with independent variables in order to assess the quality of ACT, RCT and λ_k by means of the estimation of Mean Squared Errors (MSE). Let λ_k^{oi} be the eigen value issued from the contamination of the data $\mathbf{X}$ by an outlier oi and λ_k the eigen value estimated without contamination. Over 1,000 different possible contaminations, the MSE of λ_k is given by:

$$MSE_{\lambda_k} = \frac{\sum_{i=1}^{1,000} (\lambda_k^{oi} - \lambda_k)^2}{1,000}, \ \forall k = 1, \ldots, h. \tag{14}$$

The MSE of ACT et RCT are computed in the same manner.

Algorithm 1: Monte Carlo Simulation

Result: Robust Gini PCA with data contamination

$\theta = 1$ [θ is the value of the outlier] and $N = 500$;

repeat

  Generate a 5-variate normal distribution $\mathbf{X} \sim \mathcal{N}$;

  Contamination: 1 observation (row) of $\mathbf{X}$ is multiplied by θ [random row localization] ;

  Compute ACT^{oi}, RCT^{oi} and λ_k^{oi} ;

until $\theta = 1,000$ [increment of 1];

return Mean squared Errors of ACT, RCT and λ_k ;

The MSE of the ACT of the 500 observations are computed for components $\mathbf{f}_1$ and $\mathbf{f}_2$, Figure 1a and 1b respectively.

Figure 1a: ACT_1 Axis 1

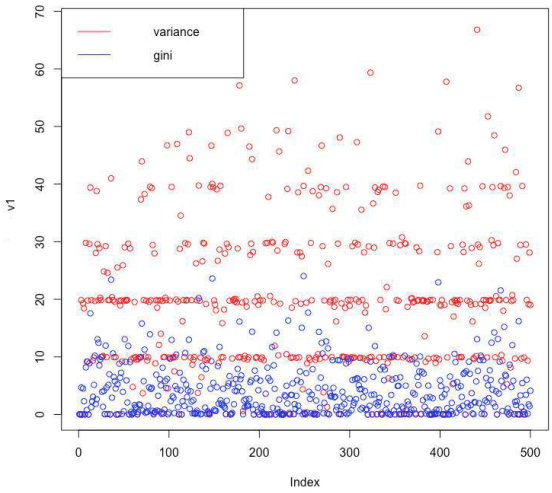


Figure 1b: ACT_2 Axis 2

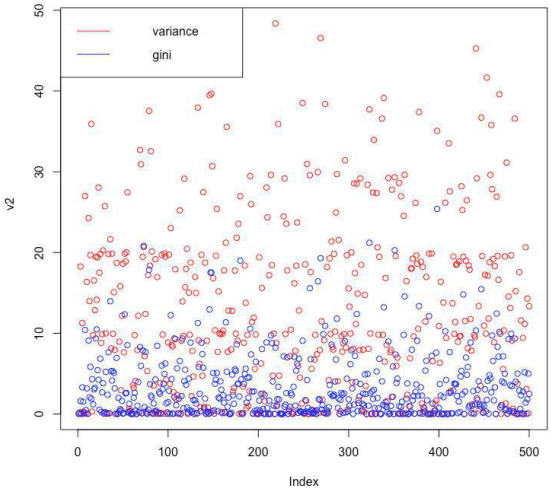


Figure 1c: RCT_1 Axis 1

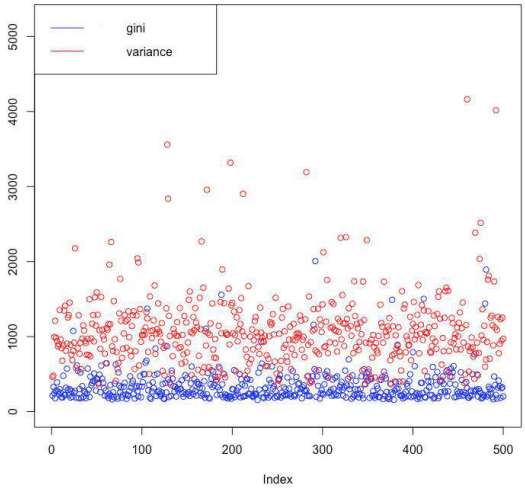
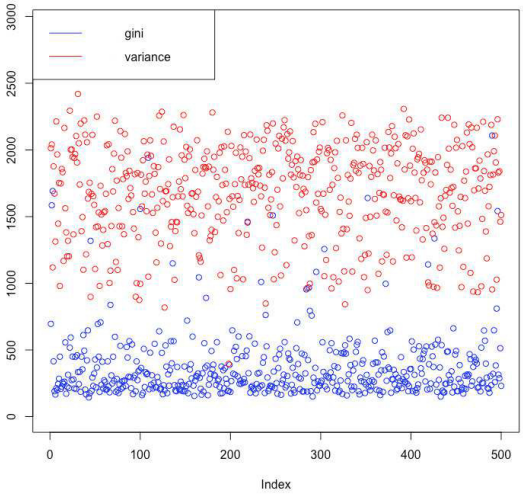


Figure 1d: RCT_2 Axis 2



The MSE issued from the Gini PCA (blue points) are less spread out than those of the variance (red points). This means that the quantity of information (dispersion) captured by each observation i remains much more stable with the Gini PCA when the data are increasingly contaminated by θ. The same conclusion holds true for the MSE of RCT (Figures 1c/1d).

Table I below depicts the MSE of the eigen values that are much lower in the Gini PCA (except on axis 3 since the quantity of dispersion is not significant on this axis). The variability on each axis (in mean over 1,000 iterations) $\frac{\lambda_k}{\sum_k \lambda_k} \times 100$ shows that the presence of one outlier drastically affects the repartition of the information on the three components. In the standard PCA (Variance case), the overall variability

is important on component 1 (90%), whereas each component must capture 1/5 of the overall variability (since the 5 variables are independent). The repartition of the variability on each component is more uniform in the Gini case.

Table I. Eigen Values and MSE: Normal distributions

	% λ_k (Gini)	% λ_k (Var)	MSE (Var)	MSE (Gini)
$axis_1$	0.57	0.90	11.89	0.69
$axis_2$	0.14	0.05	0.79	0.81
$axis_3$	0.11	0.02	0.75	14.15

Another Monte Carlo simulation is performed with a mixture of probability distributions of size $N = 500$: Normal [$\mathcal{N}$(0,1)], Gamma [Γ(2,2)], Uniform [$U(\min = 1, \max = 5)$], Cauchy [(location = 0, scale = 1)] and Beta [$\beta(2, 3)$]. The results are similar.

Figure 2a: ACT_1 Axis 1

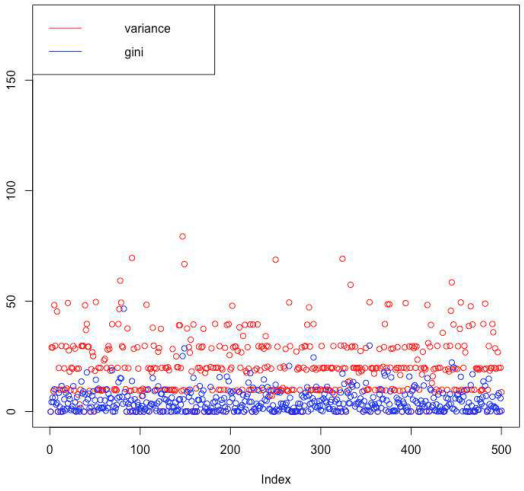


Figure 2b: ACT_2 Axis 2

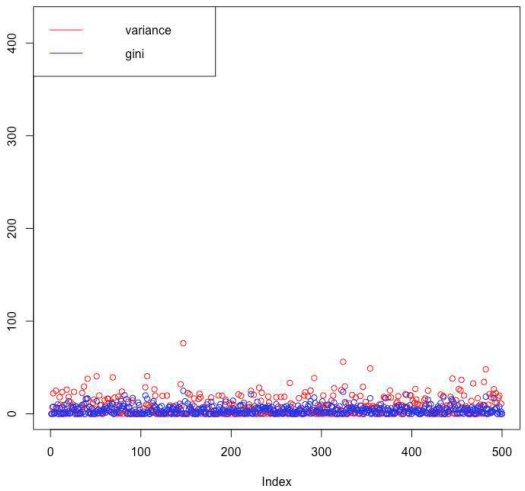


Figure 2c: RCT_1 Axis 1

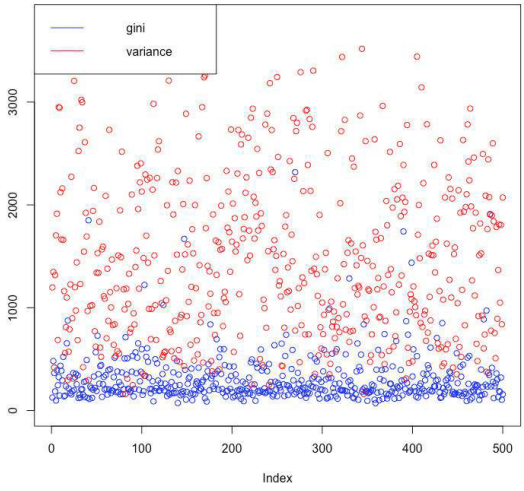


Figure 2d: RCT_2 Axis 2

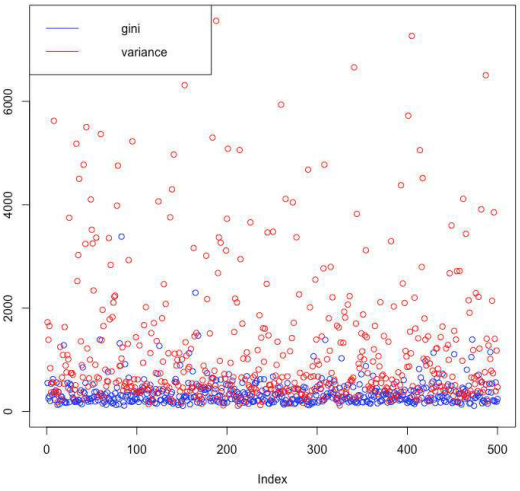


Table II. Eigen Values and MSE: mixture of distributions

	% λ_k (Gini)	% λ_k (Var)	MSE (Var)	MSE (Gini)
$axis_1$	0.64	0.88	11.04	0.49
$axis_2$	0.14	0.09	0.87	0.81
$axis_3$	0.09	0.01	0.74	19.25

4 - Concluding remarks

In this paper, a robust Gini PCA has been performed thanks to the cogini operator underlying the Gini correlation matrix $GC(\mathbf{X})$. The interpretations of the Gini PCA, on the basis of ACT, RCT and eigen values, have been shown to be more relevant than the variance case when one outlier affects the sample. This opens the way on using the Gini PCA in many fields. For instance, in financial econometrics, the principal axes are employed as risk factors in order to compute systematic risks and to deduce the risk premium. This also open the way on generalized Gini PCA, which could be based on the generalized cogini operator, see Yitzhaki and Schechtman (2013) and Greselin and Zitikis (2015).

References

A. Baccini, P. Besse and A. de Falguerolles (1996), " A L_1 norm PCA and a heuristic approach ", in *Ordinal and Symbolic Data Analysis*, E Didday, Y. Lechevalier and O. Opitz (eds), Springer, 359-368.

Carcea, M. and R. Serfling (2015), " A Gini Autocovariance Function for Time Series Modelling ", *Journal of Times Series Analysis* **36(6)**, 817-838.

Ding, C., Zhou, D., Ha, X., Zha, H. (2006), " R_1-PCA: Rotational Invariant L_1-norm Principal Component Analysis for Robust Subspace Factorization ", Proceedings of the 23 rd International Conference on Machine Learning, Pittsburgh.

Gini C. (1912), " Variabilità e mutabilità, Memori di Metodologia Statistica ", Vol. 1, Variabilità e Concentrazione. Libreria Eredi Virgilio Veschi, Rome, 211-382.

Greselin, F. and R. Zitikis (2015), " Measuring Economic Inequality and Risk: A Unifying Approach Based on Personal Gambles, Societal Preferences and References ", *Electronic SSRN Journal*.

Olkin, I. and S. Yitzhaki (1992), " Gini Regression Analysis ", *International Statistical Review* **60(2)**, 185-196.

Schechtman, E. and S. Yitzhaki (1987), " A Measure of Association Based on Gini's Mean Difference ", *Communications in Statistics* **A16**, 207-231.

Schechtman, E. and S. Yitzhaki (2003), " A family of correlation coefficients based on the extended Gini index ", *Journal of Economic Inequality* **1(2)**, 129-146.

Yitzhaki, S. (2003)," Gini's Mean difference: a superior measure of variability for non-normal distributions ", *Metron* **LXI(2)**, 285-316.

Yitzhaki, S. and P. Lambert (2013), " The Relationship Between the Absolute Deviation from a Quantile and Gini's Mean Difference " *Metron* **71**, 97-104.

Yitzhaki, S. and E. Schechtman (2013), *The Gini Methodology. A Primer on a Statistical Methodology*, Springer.