Lorenz model selection

January 7, 2020

Abstract

In the paper we introduce novel model selection measures based on Lorenz Zonoids which, differently from measures based on correlations, are based on a mutual notion of variability and are more robust to the presence of outlying observations. By means of Lorenz Zonoids, which in the univariate case correspond to the Gini coefficient, the contribution of each explanatory variable to the predictive power of a linear model can be measured more accurately. Exploiting Lorenz Zonoids, we develop a Marginal Gini Contribution measure that allows to measure the absolute explanatory power of any covariate, and a Partial Gini Contribution measure that allows to measure the additional contribution of a new covariate to an existing model.

Keywords: Dependence measures, Linear models, Lorenz Zonoids, Marginal Gini Contribution, Partial Gini Contribution.

1 Introduction

A very important problem in statistics and in data analysis is to compare alternative models on a given set of data, for example in terms of their predictive accuracy. The traditional paradigm compares statistical models through a sequence of pairwise comparisons, which eventually leads to a statistical test, that provides a threshold which can be used to decide which model to adopt. Statistical model comparison is, however, generally not applicable to machine learning models, which do not necessarily have an underlying probabilistic model. In this case, models are compared in terms of information criteria such as AIC or BIC which, while providing a total ordering of models, require thresholds to choose among them (Hand, Mannilla and Smyth, Chapter 11, 2001; Burnham and Anderson, 2004).

To overcome this problem, the last few years have witnessed the growing importance of model comparison methods based on the direct calculation of the predictive accuracy of a model, through cross-validation methods. In the cross-validation process, the dataset is split in two or more datasets, with training datasets used to fit a model and validation datasets used to compare the predictions obtained with the fitted model on the validation dataset. When the response variable is continuous, a typical cross-validation summary criterion is the root mean squared error (RMSE) which calculates the difference between the observed and the predicted values. This difference can be compared with a threshold value to choose among competing models (see e.g. Diebold and Mariano 1995).

A problem with cross-validation measures, such as the RMSE, is that they are not normalised, similarly to information criteria, differently from what occurs in statistical model comparison. A further problem is that, when the number of explanatory variables increases, the RMSE does not necessarily decrease.

We aim to overcome these drawbacks with a new model comparison dependence measure

that is normalised, like statistical tests, but can also be applied to machine learning models.

The most commonly used measure to detect a relation of dependence between a response variable and a set of explanatory variables is the coefficient of determination R^2. The coefficient of determination, although widely employed, has some drawbacks. For example, it is based on the distance between each observation and the mean point and, consequently, may be affected by extreme observations (Rousseeuw and Leroy, Chapters 1-2, 1987). We propose to overcome this problem with the definition of a new measure of dependence, based on Lorenz Zonoids. To do so, we extend the work of Giudici and Raffinetti (2011), who introduced a decomposition of the classical Gini coefficient in terms of concordance and discordance shares.

The new measure is normalized and enjoys an "inclusion property" which leads to values that increase with the number of explanatory variables, similarly to the R^2, but differently from the RMSE. In addition, it is based on the mutual distance between the observations, rather than on the distance from the mean value and, therefore, is less affected by extreme observations.

The rest of the paper is organized as follows. Section 2 provides a background on Lorenz Zonoids, especially on its main features and properties. Section 3 introduces our proposed Lorenz Zonoid dependence measures, in the linear model framework. To better understand our proposal, Section 4 includes an illustrative example and a real application to bitcoin price discovery. Finally, Section 5 briefly concludes the paper.

2 Background

The Lorenz Zonoid has been introduced by Koshevoy (1995) for empirical distributions and by Mosler (1994) for general probability distributions. The Lorenz Zonoid of a d-dimensional random vector corresponds to a convex set in $\mathbb{R}^{d+1}$, whose role is to analyse and compare random vectors. Through the Lorenz Zonoid representation one can establish an ordering of random vectors that reflects their variability: the investigation of such ordering is induced by the inclusion between subsequent Lorenz Zonoids. This aspect provides a helpful support for our proposed development.

We now define the Lorenz curve for a non-negative variable Y, following Koshevoy and Mosler (2007). The Lorenz curve of a non negative random variable Y having expectation $E(Y) = \mu$ is the graph of the function

$$t \mapsto \mu^{-1} \int_0^t F_Y^{-1}(s)ds, 0 \leq t \leq 1,$$

where F_Y^{-1} is the quantile function of Y: $F_Y^{-1} = \min\{y : F(y) \geq t\}$.

Roughly speaking, given n observations, the Lorenz Curve L_Y of the Y variable (see Lorenz 1905) is given by the set of points $(i/n, \sum_{j=1}^{i} y_{(j)}/(n\bar{y}))$, for $i = 1, \ldots, n$, where $y_{(i)}$ indicates the Y variable values ordered in a non-decreasing sense and $\bar{y}$ is the Y variable mean value. Analogously, the Y variable can be re-ordered in a non-increasing sense providing the dual Lorenz curve $L_Y^{'}$, which is defined as the set of points $(i/n, \sum_{j=1}^{i} y_{(n+1-j)}/(n\bar{y}))$. The area lying between the L_Y and $L_Y^{'}$ Lorenz curves corresponds to the Gini coefficient, which is typically employed as an indicator of inequality, especially when dealing with income data.

When considering more than one variable, the generalisation of the Lorenz curve in d

dimensions is the so-called Lorenz Zonoid.

The Lorenz Zonoid of a general d-variate random vector can be defined following Koshevoy and Mosler (1996). Consider a set $\mathcal{Y}^{d+}$ of random vectors in $\mathbb{R}^d$ that have finite and positive (in each component) expectation and, within this set, the subset $\mathcal{Y}_+^{d+} \subset \mathcal{Y}^{d+}$ of those vectors that have support in $\mathbb{R}_+^d$.

For $\mathbf{Y} \in \mathcal{Y}^{d+}$, we introduce the notation

$$\tilde{\mathbf{Y}} = \left(\frac{Y_1}{E(Y_1)}, \ldots, \frac{Y_d}{E(Y_d)} \right),$$

which is the vector component wise divided by its expectation.

The Lorenz Zonoid of a random vector $\mathbf{Y} \in \mathcal{Y}^{d+}$ is a convex compact set in $\mathbb{R}^{d+1}$, defined as follows:

$$LZ(\mathbf{Y}) = \left\{ E[(g(\tilde{\mathbf{Y}}), g(\tilde{\mathbf{Y}})\tilde{\mathbf{Y}})] : g : \mathbb{R}^d \to [0,1] \text{ measurable} \right\}.$$

For the sake of clarity, a function $g : E \to \mathbb{R}$ is measurable if E is a measurable set and for each real number $r \in \mathbb{R}$, the set $\{y \in E : g(y) > r\}$ is measurable. It follows that continuous and monotone functions are measurable. We remark that if $\mathbf{X} \in \mathcal{Y}_+^{d+}$, i.e. has support in $\mathbb{R}_+^d$, the Lorenz Zonoid is contained in the hypercube of $\mathbb{R}^{d+1}$.

The Lorenz Zonoid fulfills many attractive properties, some of which are the building blocks for the contribution proposed here.

***Property* 1** *The Lorenz Zonoid induces a linear dependence order:*

$$\mathbf{Y} \preceq_{ld} \mathbf{X} \text{ if } LZ(\mathbf{X}) \subset LZ(\mathbf{Y}), \tag{1}$$

where $LZ(\mathbf{X})$ and $LZ(\mathbf{Y})$ are the Lorenz Zonoids of the random vectors $\mathbf{X}$ and $\mathbf{Y}$ and where $\preceq_{ld}$ indicates a linear dependence order (see, for instance, Dall'Aglio and Scarisni 2003).

***Property* 2** *The Lorenz Zonoid induces a dominance order:*

$$\mathbf{Y} \preceq_L \mathbf{X} \text{ if } LZ(\mathbf{X}) \subset LZ(\mathbf{Y}),$$

where $\preceq_L$ defines Lorenz dominance (see, for instance, Koshevoy and Mosler 2007):

An important corollary of property 2 is that, in the univariate case, there is a perfect equivalence between the Lorenz Zonoid order and the order induced by the variability order:

***Corollary* 1** $\mathbf{Y} \preceq_{dil} \mathbf{X} \Leftrightarrow \mathbf{Y} \preceq_L \mathbf{X}$,

where $\preceq_{dil}$ indicates that $Var(Y) \geq Var(X)$. In other words, through the Lorenz dominance an ordering based on the variability can be equivalently specified.

Within the univariate context, let us denote the Gini coefficient with the notation $LZ_{d=1}(\cdot)$, to indicate the Lorenz Zonoid in the univariate case. The condition of linear dependence reported in Property 1, can be further re-formalized to cover the case of variables whose linear dependence may be investigated through a linear regression model.

***Proposition* 1** *Consider the bidimensional vector (Y, X) and apply a linear regression model, such that $\hat{y} = \hat{\alpha} + \hat{\beta}x$. Assume that $\hat{Y}$ takes non-negative values. Denote respectively with $L_Y(t)$ and $L'_Y(t)$ the Y Lorenz curve and its dual, and with $L_{\hat{Y}}(t)$ and $L'_{\hat{Y}}(t)$ the $\hat{Y}$ Lorenz curve and its dual. One can prove (see e.g. Muliere and Petrone 1992) that $L_Y(t) \leq L_{\hat{Y}}(t)$ where $L'_Y(t) = \frac{1}{E(Y)} \int_{1-t}^{1} F_Y^{-1}(s)ds$, $0 \leq t \leq 1$. Furthermore, $L'_{\hat{Y}}(t) \leq L'_Y(t)$.*

Proposition 1 provides a very important "*inclusion property*" which parallels what occurs to the variance explained by the regression: $Var(\hat{Y}) \leq Var(Y)$.

In other words, the existence of a linear dependence relationship between Y and X translates into an inclusion between the response variable Y and the linear estimated variable $\hat{Y}$ Lorenz Zonoids. Figure 1 shows this outcome in a pictorial way.

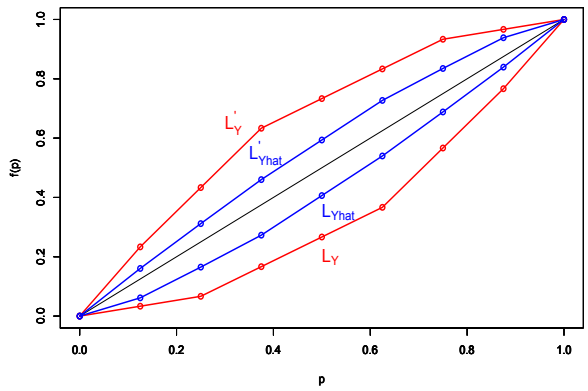

Figure 1. Visualization of the Y Lorenz Zonoid (area between red lines) and $\hat{Y}$ Lorenz Zonoid (area between blue lines).

Figure 1 shows that the $\hat{Y}$ Lorenz Zonoid, based on the estimates obtained from the linear regression of Y on X, is contained in the Y variable Lorenz Zonoid.

3 Proposal

We exploit the Lorenz Zonoid ($LZ_{d=1}(\cdot)$) as a measure of variability that characterizes a phenomenon of interest. While the variance measures the variability with repect to the mean, the Lorenz Zonoid measures the mutual variability. Similarly to the variance, the

Lorenz Zonoid can be used within linear models to assess the contribution of additional independent variables in explaining the variability of the response variable.

This is the aim of our proposal: to introduce a new dependence measure, addressed to explain the response variable Lorenz Zonoid share "explained" by any additional independent variable, and to generalize this into a stepwise model selection procedure based on the Lorenz Zonoid explained shares. To our knowledge, this is the only contribution available, apart from a short and purely theoretical communication by Raffinetti and Giudici (2013).

Let $LZ_{d=1}(Y)$ be the Lorenz Zonoid of the response variable Y, and X_1 an independent variable such that $\hat{Y}_{X_1}$ is the vector of the estimated values computed with a linear regression model such that $\hat{Y}_{X_1} = \hat{\alpha} + \hat{\beta}X_1$. Define with $LZ_{d=1}(\hat{Y}_{X_1})$ the Lorenz Zonoid of $\hat{Y}_{X_1}$. Consider an additional independent variable X_2 and a corresponding linear regression model such that $\hat{Y}_{X_2} = \hat{\alpha} + \hat{\beta}X_2$. Define with $LZ_{d=1}(\hat{Y}_{X_2})$ the corresponding Lorenz Zonoid.

A very useful result, contained in Lerman and Yitzhaki (1984) is that, in the univariate case, the Lorenz Zonoid of a variable may be expressed by resorting to the covariance operator. Formally we have

$$LZ_{d=1}(Y) = \frac{2Cov(Y, F(Y))}{\mu}, \tag{2}$$

where μ is the response variable Y mean value and $F(Y)$ is the distribution function of Y. In the same manner, $LZ_{d=1}(\hat{Y}_{X_1})$ and $LZ_{d=1}(\hat{Y}_{X_2})$ can be expressed as

$$LZ_{d=1}(\hat{Y}_{X_1}) = \frac{2Cov(\hat{Y}_{X_1}, F(\hat{Y}_{X_1}))}{\mu} \text{ and } LZ_{d=1}(\hat{Y}_{X_2}) = \frac{2Cov(\hat{Y}_{X_2}, F(\hat{Y}_{X_2}))}{\mu}, \tag{3}$$

where $E(\hat{Y}_{X_1}) = E(E(Y|\hat{Y}_{X_1})) = \mu$ and $E(\hat{Y}_{X_2}) = E(E(Y|\hat{Y}_{X_2})) = \mu$, $F(\hat{Y}_{X_1})$ and $F(\hat{Y}_{X_2})$ are the distribution functions of $\hat{Y}_{X_1}$ and $\hat{Y}_{X_2}$, respectively.

It can be shown that (2) and (3) can be equivalently expressed in term of rank scores. The following holds.

Result 1. Let $r(Y)$, $r(\hat{Y}_{X_1})$ and $r(\hat{Y}_{X_2})$ be the rank scores corresponding to the Y, $\hat{Y}_{X_1}$ and $\hat{Y}_{X_2}$ variables. Since the $r(\cdot)$ terms are the empirical representation of $F(\cdot) = r(\cdot)/n$, it can then be shown that:

$$LZ_{d=1}(Y) = \frac{2Cov(Y, r(Y))}{n\mu}, LZ_{d=1}(\hat{Y}_{X_1}) = \frac{2Cov(\hat{Y}_{X_1}, r(\hat{Y}_{X_1}))}{n\mu}$$
$$\text{and } LZ_{d=1}(\hat{Y}_{X_2}) = \frac{2Cov(\hat{Y}_{X_2}, r(\hat{Y}_{X_2}))}{n\mu}. \tag{4}$$

Proof. Consider the response variable Y. We have to prove that

$$LZ_{d=1}(Y) = \frac{2Cov(Y, F(Y))}{\mu} = \frac{2Cov(Y, r(Y))}{n\mu}. \tag{5}$$

The term $Cov(Y, F(Y))$ is equivalent to $Cov\left(Y, \frac{r(Y)}{n}\right)$. Through some computations, we obtain that

$$Cov\left(Y, \frac{r(Y)}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} Y_i \frac{r(Y_i)}{n} - \mu\frac{\bar{r}(Y)}{n} = \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i r(Y_i) - \mu\bar{r}(Y)\right]$$
$$= \frac{1}{n}Cov(Y, r(Y)) = \frac{1}{n}Corr(Y, r(Y))\sigma_Y\sigma_{r(Y)}$$

where $\bar{r}(Y)$ is the mean of $r(Y)$, σ_Y and $\sigma_{r(Y)}$ are the standard deviations, respectively, of Y and $r(Y)$. The equivalence in (5) follows.

From an interpretational viewpoint Result 1. shows that the Lorenz Zonoid is proportional to the correlation between the response variable and its ranks. We recall that the variance is proportional to the squared distance of the response variable from its mean.

The previous result easily generalizes to $\hat{Y}_{X_1}$ and $\hat{Y}_{X_2}$ and, therefore, the Lorenz Zonoid share explained by a linear model is proportional to the correlation between the fitted values and their ranks.

Thus, when the employed measure of variability is the Lorenz Zonoid, the goodness of fit of a regression is proportional to the correlation between the fitted values and their ranks. In contrast, in the standard linear model case, the goodness of fit of a regression is proportional to the squared distance between the fitted values and the mean.

***Remark* 1** *Given a sample data of size n, the formulas in (2) and (3) may be re-expressed as:*

$$LZ_{d=1}(y) = \frac{2Cov(y, r(y))}{n\bar{y}}, LZ_{d=1}(\hat{y}_{x_1}) = \frac{2Cov(\hat{y}_{x_1}, r(\hat{y}_{x_1}))}{n\bar{y}}$$
$$and\ LZ_{d=1}(\hat{y}_{x_2}) = \frac{2Cov(\hat{y}_{x_2}, r(\hat{y}_{x_2}))}{n\bar{y}} \tag{6}$$

where y, $\hat{y}_{x_1}$ and $\hat{y}_{x_1}$ are the vectors of the observed and estimated values, $r(y)$, $r(\hat{y}_{x_1})$ and $r(\hat{y}_{x_2})$ are the ranks of the observed values, and $\bar{y}$ is the sample mean.

It can also be shown that the Lorenz Zonoid can be expressed as a function of the sum of the distances between the y-axis values of the points lying on the Lorenz curve and those of the points lying on the bisector curve (the black curve in Figure 1). To be able to show this, we first need to derive the expression of the distance.

Result 2. Let Y be a response variable, whose values arranged in non-decreasing sense are denoted with $y_{(i)}$, for $i = 1, \ldots, n$. Let q be the sum of the distances between the y-axis values of the points lying on the Lorenz curve and those of the points lying on the bisector curve. It can then be shown that

$$q = \frac{1}{\bar{y}} cov(y_{(i)}, r(y)). \tag{7}$$

Proof. Consider the coordinates $(i/n, \sum_{j=1}^{i} y_{(j)}/n\bar{y})$ of the points lying on the Lorenz curve of the response variable Y and the coordinates $(i/n, i/n)$ of the points lying on the bisector curve. It follows that q can be defined as

$$q = \sum_{i=1}^{n} \left\{ \frac{i}{n} - \frac{1}{n\bar{y}} \sum_{j=1}^{i} y_{(j)} \right\} = \sum_{i=1}^{n} \frac{i}{n} - \frac{1}{n\bar{y}} \sum_{i=1}^{n} \sum_{j=1}^{i} y_{(j)}. \tag{8}$$

Because $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$ and $\sum_{i=1}^{n} \sum_{j=1}^{i} y_{(j)} = n(n+1)\bar{y} - \sum_{i=1}^{n} i y_{(i)}$, the term on the right hand side of equation (8) can be written as

$$\begin{aligned} q = \frac{n(n+1)}{2n} - \frac{1}{n\bar{y}} \left[n(n+1)\bar{y} - \sum_{i=1}^{n} i y_{(i)} \right] &= \frac{1}{\bar{y}} \left[\frac{1}{n} \sum_{i=1}^{n} i y_{(i)} - \frac{n(n+1)}{2n} \bar{y} \right] \\ &= \frac{1}{\bar{y}} \left[\frac{1}{n} \sum_{i=1}^{n} i y_{(i)} - \frac{(n+1)}{2} \bar{y} \right]. \end{aligned} \tag{9}$$

The term within the square brackets in (9) corresponds to the covariance between the Y values and their ranks, where the mean of i (i.e., $r(y)$) is equal to $\bar{i} = (n+1)/2$. Then, the equivalence in (7) follows.

We are now able to show that the Lorenz Zonoid is a function of the sum of the distances between the y-axis values of the points lying on the Lorenz curve and those of the points lying on the bisector curve. The following result demonstrates the equivalence.

Result 3. From equations (7) and (9), it follows that

$$LZ_{d=1}(y) = \frac{2}{n} q = \frac{2}{n\bar{y}} \left[\frac{1}{n} \sum_{i=1}^{n} i y_{(i)} - \frac{n(n+1)}{2n} \bar{y} \right]. \tag{10}$$

The previous result easily generalizes also for variables $\hat{Y}_{X_1}$ and $\hat{Y}_{X_2}$, which are the estimated fitted values. Let $\hat{y}_{(x_1 i)}$ and $\hat{y}_{(x_2 i)}$, for $i = 1, \ldots, n$, be the $\hat{Y}_{X_1}$ and $\hat{Y}_{X_2}$ values arranged in a non-decreasing sense. Similarly to (10), $LZ_{d=1}(\hat{y}_{x_1})$ and $LZ_{d=1}(\hat{y}_{x_2})$ may be re-expressed as

$$LZ_{d=1}(\hat{y}_{x_1}) = \frac{2}{n\bar{y}}\left[\frac{1}{n}\sum_{i=1}^{n} i\hat{y}_{(x_1 i)} - \frac{n(n+1)}{2n}\bar{y}\right] \tag{11}$$

$$LZ_{d=1}(\hat{y}_{x_2}) = \frac{2}{n\bar{y}}\left[\frac{1}{n}\sum_{i=1}^{n} i\hat{y}_{(x_2 i)} - \frac{n(n+1)}{2n}\bar{y}\right]. \tag{12}$$

We can now employ the previous results to derive a marginal dependence measure, which will be denoted by MGC (*Marginal Gini Contribution*). The measure can evaluate the Y Lorenz Zonoid share marginally explained by a single explanatory variable X_h for $h = 1, \ldots, k$. It can be defined as

$$MGC_{(Y|X_h)} = \frac{LZ_{d=1}(\hat{Y}_{X_h})}{LZ_{d=1}(Y)} = \frac{2Cov(\hat{Y}_{X_h}, r(\hat{Y}_{X_h}))/n\mu}{2Cov(Y, r(Y))/n\mu} = \frac{Cov(\hat{Y}_{X_h}, r(\hat{Y}_{X_h}))}{Cov(Y, r(Y))}, \tag{13}$$

whose sample version is

$$MGC_{(y|x_h)} = \frac{\frac{2}{n\bar{y}}\left[\frac{1}{n}\sum_{i=1}^{n} i\hat{y}_{(x_h i)} - \frac{n(n+1)}{2n}\bar{y}\right]}{\frac{2}{n\bar{y}}\left[\frac{1}{n}\sum_{i=1}^{n} iy_{(i)} - \frac{n(n+1)}{2n}\bar{y}\right]} = \frac{Cov(\hat{y}_{x_h}, r(\hat{y}_{x_h}))}{Cov(y, r(y))}. \tag{14}$$

The MGC measure may be used to select the explanatory variables in a regression context. For example, the explanatory variable with the largest contribution in explaining the share of the response variable Lorenz Zonoid measured by the MGC can be chosen as an explanatory variable in a regression model.

To understand whether further variables can improve a given regression model, we need to define a partial contribution measure. This can be done extending the MGC definition into a model selection procedure.

In the general context, characterized by k explanatory variables, we would like to determine the effect related to the introduction of a new $(k+1)$-th explanatory variable into a linear regression model. The inclusion of a new explanatory variable provides an enlargement of the $\hat{Y}$ Lorenz Zonoid. The Lorenz Zonoid of the Y linear estimated values, denoted with $LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_k})$, corresponds to the dilation measure of the Y response variable Lorenz Zonoid $LZ_{d=1}(Y)$. Therefore, the introduction of an additional covariate in multiple linear regression models translates into an increase of the "explained" Y variability.

In the well-known linear regression model, the contribution of a single variable to the regression plane is additive and, therefore, the addition of a new explanatory variable translates into an increase of the multiple determination coefficient (see e.g. Giudici, Chapter 4, 2003). More precisely, suppose a linear regression model is built that is characterized by k explanatory variables. Let us introduce an additional $(k+1)$-th explanatory variable. Its contribution determines an increase of the Y variable "explained" variability, defined as the difference between $Var(\hat{Y}_{X_1,\ldots,X_{k+1}})$ and $Var(\hat{Y}_{X_1,\ldots,X_k})$, where $Var(\hat{Y}_{X_1,\ldots,X_k})$ denotes the Y variability "explained" by the $X_1,\ldots,X_k$ independent variables whereas $Var(\hat{Y}_{X_1,\ldots,X_{k+1}})$ denotes the Y variability "explained" by the $X_1,\ldots,X_{k+1}$ independent variables. The squared partial correlation coefficient is expressed as

$$r^2_{Y,X_{k+1}|X_1,\ldots,X_k} = \frac{Var(\hat{Y}_{X_1,\ldots,X_{k+1}}) - Var(\hat{Y}_{X_1,\ldots,X_k})}{Var(Y) - Var(\hat{Y}_{X_1,\ldots,X_k})}, \tag{15}$$

where $Var(Y) - Var(\hat{Y}_{X_1,\ldots,X_k})$ identifies the Y variable variability not explained by the

$X_1, \ldots, X_k$ covariates.

We aim at building a partial contribution measure that "parallels" the partial correlation coefficient construction. Specifically, we propose as partial contribution measure the ratio between a numerator characterized by a term denoting the contribution generated by the $(k+1)$-*th* explanatory variable and a denominator including a term which describes the share of the Y Lorenz Zonoid "not explained" by the $\hat{Y}_{X_k}$ Lorenz Zonoid. The additional contribution related to the $(k+1)$-*th* explanatory variable inclusion can be measured through the difference between the $\hat{Y}_{X_1,\ldots,X_{k+1}}$ and $\hat{Y}_{X_1,\ldots,X_k}$ Lorenz Zonoids, that is $LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_{k+1}}) - LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_k})$.

A relative index, measuring the additional contribution provided by the X_{k+1} independent variable can then be obtained in analogy with the partial correlation coefficient construction. Such a measure, which we will call *PGC* (*Partial Gini Contribution*), can be expressed as:

$$PGC_{Y,X_{k+1}|X_1,\ldots,X_k} = \frac{LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_{k+1}}) - LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_k})}{LZ_{d=1}(Y) - LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_k})}. \tag{16}$$

Note that equation (16) can be re-expressed, in analogy with the partial correlation coefficient, in terms of covariances:

$$\begin{aligned} PGC_{Y,X_{k+1}|X_1,\ldots,X_k} &= \frac{\frac{2}{n\mu}Cov(\hat{Y}_{X_1,\ldots,X_{k+1}}, r(\hat{Y}_{X_1,\ldots,X_{k+1}})) - \frac{2}{n\mu}Cov(\hat{Y}_{X_1,\ldots,X_k}, r(\hat{Y}_{X_1,\ldots,X_k}))}{\frac{2}{n\mu}Cov(Y, r(Y)) - \frac{2}{n\mu}Cov(\hat{Y}_{X_1,\ldots,X_k}, r(\hat{Y}_{X_1,\ldots,X_k}))} \\ &= \frac{Cov(\hat{Y}_{X_1,\ldots,X_{k+1}}, r(\hat{Y}_{X_1,\ldots,X_{k+1}})) - Cov(\hat{Y}_{X_1,\ldots,X_k}, r(\hat{Y}_{X_1,\ldots,X_k}))}{Cov(Y, r(Y)) - Cov(\hat{Y}_{X_1,\ldots,X_k}, r(\hat{Y}_{X_1,\ldots,X_k}))}. \end{aligned} \tag{17}$$

We remark that, for the first variable (denoted h) included in the model, the equivalence $MGC_{(Y|X_h)} = PGC_{Y|X_h}$ holds.

Result 4. It can be shown that, after some manipulations, $PGC_{Y,X_{k+1}|X_1,\ldots,X_k}$ computed on sample data can be expressed as:

$$PGC_{y,x_{k+1}|x_1,\ldots,x_k} = \frac{\sum_{i=1}^{n} i(\hat{y}_{(x_1,\ldots,x_{k+1}i)} - \hat{y}_{(x_1,\ldots,x_ki)})}{\sum_{i=1}^{n} i(y_{(i)} - \hat{y}_{(x_1,\ldots,x_ki)})}. \tag{18}$$

***Property* 3** *Under the condition of linear dependence between Y and the k explanatory variables, it holds that $0 \leq PGC_{Y,X_{k+1}|X_1,\ldots,X_k} \leq 1$. In the intermediate scenarios the PGC measure takes values always smaller than 1 or greater than 0, depending on the contribution provided by the additional X_{k+1} covariate to the explanation of the response variable.*

Proof. The following inequalities have to be proven:

a) $\frac{LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_{k+1}}) - LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_k})}{LZ_{d=1}(Y) - LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_k})} \geq 0$;

b) $\frac{LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_{k+1}}) - LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_k})}{LZ_{d=1}(Y) - LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_k})} \leq 1$.

It is worth noting that the denominator of (16) is always positive. The only case in which $LZ_{d=1}(Y) - LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_k}) = 0$ is reached, is if the k covariates perfectly explain the response variable. In this case, no additional explanatory variable needs to be considered for inclusion in the model. Moreover, no negative values are allowed due to the inclusion of $LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_k})$ into $LZ_{d=1}(Y)$. From inequality a) it follows that

$$LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_{k+1}}) - LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_k}) \geq 0 \Rightarrow LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_{k+1}}) \geq LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_k}). \tag{19}$$

The relation in (19) is always fulfilled since the inclusion of a new explanatory variable into the model typically provides an enlargement of the Lorenz Zonoid built on the corresponding linear estimated values. Only in the case that the additional explanatory variable does not provide any improvement in the explained variability of the response variable, it results that $LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_{k+1}}) - LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_k}) = 0$, which is equal to

$LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_{k+1}}) = LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_k})$. The same conclusion may be obtained by resorting to formulas (17), in terms of covariances, and (18), when referring to the sample data. Specifically, it follows that $Cov(\hat{Y}_{X_1,\ldots,X_{k+1}}, r(\hat{Y}_{X_1,\ldots,X_{k+1}})) \geq Cov(\hat{Y}_{X_1,\ldots,X_k}, r(\hat{Y}_{X_1,\ldots,X_k}))$ and $\hat{y}_{(x_1,\ldots,x_{k+1}i)} \geq \hat{y}_{(x_1,\ldots,x_k i)}$.

Next, from inequality b) it follows that

$$LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_{k+1}}) \leq LZ_{d=1}(Y). \tag{20}$$

The condition in (20) is a direct consequence of the inclusion property. Indeed the Y variability explained by the $X_1, \ldots, X_{k+1}$ covariates, here defined in terms of the Lorenz Zonoid built on the linear estimated values provided by the linear regression model, corresponds at most to the total variability underlying theresponse variable. Equality between $LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_{k+1}})$ and $LZ_{d=1}(Y)$ is achieved if the linear model built on the $X_1, \ldots, X_{k+1}$ covariates perfectly explains the target variable Y. as for inequality a), the relation in (20) can be expressed in terms of covariances and in terms of sample data as $Cov(\hat{Y}_{X_1,\ldots,X_{k+1}}, r(\hat{Y}_{X_1,\ldots,X_{k+1}})) \leq Cov(Y, r(\hat{Y}))$ and $\hat{y}_{(x_1,\ldots,x_k)i} \leq y_{(i)}$.

In the standard linear model selection context, it is well known that the multiple coefficient of determination is related to the partial correlation coefficient of each explanatory variable. We would like a similar relationship to hold for Lorenz Zonoids as well. The following can be proved.

Result 5. In the standard model selection context with k explanatory variables, one can prove that the overall contribution of the fitted plane depends on the single contributions through the following recursive relationship:

$$R^2_{Y,X_1,\ldots,X_k} = \sum_{j=1}^{k} r^2_{Y,X_j|X_{i<j}}(1 - R^2_{Y,X_1,\ldots,X_{j-1}}), \tag{21}$$

where $R^2_{Y,X_1,\ldots,X_k}$ represents the determination coefficient of the model built on the k explanatory variables, $R^2_{Y,X_1,\ldots,X_{j-1}}$ denotes the coefficient of multiple correlation between Y and the fitted plane determined by the explanatory variables $X_1,\ldots,X_{j-1}$, and $r_{Y,X_j|X_{i<j}}$ denotes the coefficient of partial correlation between Y and X_j, conditional on the explanatory variables previously included into the model (see, e.g. Giudici, Chapter 4, 2003). We now show that the overall contribution provided by the k covariates to the explanation of the non-negative response variable Y mutual variability depends on the single contribution measured in terms of the PGC measures according to the following recursive relationship:

$$MGC_{(Y|X_1,\ldots,X_k)} = \sum_{j=1}^{k} PGC_{Y,X_j|X_{i<j}}(1 - MGC_{(Y|X_1,\ldots,X_{j-1})}), \tag{22}$$

where $MGC_{(Y|X_1,\ldots,X_k)}$ denotes the overall Y mutual variability explained by all the involved variables (i.e., $LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_k})$), $PGC_{Y,X_j|X_{i<j}}$ is the contribution associated with the j-th explanatory variable included in the model and $MGC_{(Y|X_1,\ldots,X_{j-1})}$ is the overall contribution provided by the $(j-1)$-th explanatory variables (i.e., $LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_{j-1}})$), with $j = 1,\ldots,k$.

Proof. The aim is to prove the equivalence in (22). For the sake of simplicity we consider the case of three explanatory variables ($k = 3$). We start by fitting first X_1, then X_2 and finally X_3. The relationship in (22) becomes:

$$\begin{aligned}
&MGC_{(Y|X_1,X_2,X_3)} = \sum_{j=1}^{3} PGC_{Y,X_j|X_{i<j}}(1 - MGC_{(Y,X_1,\ldots,X_{j-1})}) \\
&\frac{LZ_{d=1}(\hat{Y}_{X_1,X_2,X_3})}{LZ_{d=1}(Y)} = \frac{LZ_{d=1}(\hat{Y}_{X_1})}{LZ_{d=1}(Y)} + \frac{LZ_{d=1}(\hat{Y}_{X_1,X_2}) - LZ_{d=1}(\hat{Y}_{X_1})}{LZ_{d=1}(Y) - LZ_{d=1}(\hat{Y}_{X_1})}\left[1 - \frac{LZ_{d=1}(\hat{Y}_{X_1})}{LZ_{d=1}(Y)}\right] \\
&+ \frac{LZ_{d=1}(\hat{Y}_{X_1,X_2,X_3}) - LZ_{d=1}(\hat{Y}_{X_1,X_2})}{LZ_{d=1}(Y) - LZ_{d=1}(\hat{Y}_{X_1,2})}\left[1 - \frac{LZ_{d=1}(\hat{Y}_{X_1,X_2})}{LZ_{d=1}(Y)}\right].
\end{aligned} \tag{23}$$

In equation (23), the term in the squared brackets corresponds to $\frac{LZ_{d=1}(Y)}{LZ_{d=1}(Y)} - \frac{LZ_{d=1}(\hat{Y}_{X_1,X_2})}{LZ_{d=1}(Y)}$,

leading to:

$$\frac{LZ_{d=1}(\hat{Y}_{X_1,X_2,X_3})}{LZ_{d=1}(Y)} = \frac{LZ_{d=1}(\hat{Y}_{X_1})}{LZ_{d=1}(Y)} + \frac{LZ_{d=1}(\hat{Y}_{X_1,X_2}) - LZ_{d=1}(\hat{Y}_{X_1})}{LZ_{d=1}(Y) - LZ_{d=1}(\hat{Y}_{X_1})}\left[\frac{LZ_{d=1}(Y) - LZ_{d=1}(\hat{Y}_{X_1})}{LZ_{d=1}(Y)}\right] + \frac{LZ_{d=1}(\hat{Y}_{X_1,X_2,X_3}) - LZ_{d=1}(\hat{Y}_{X_1,X_2})}{LZ_{d=1}(Y) - LZ_{d=1}(\hat{Y}_{X_1,2})}\left[\frac{LZ_{d=1}(Y) - LZ_{d=1}(\hat{Y}_{X_1,X_2})}{LZ_{d=1}(Y)}\right]. \tag{24}$$

Through some mathematical manipulations, (24) can be re-written as

$$\frac{LZ_{d=1}(\hat{Y}_{X_1,X_2,X_3})}{LZ_{d=1}(Y)} = \frac{LZ_{d=1}(\hat{Y}_{X_1})}{LZ_{d=1}(Y)} + \frac{LZ_{d=1}(\hat{Y}_{X_1,X_2}) - LZ_{d=1}(\hat{Y}_{X_1})}{LZ_{d=1}(Y)} + \frac{LZ_{d=1}(\hat{Y}_{X_1,X_2,X_3}) - LZ_{d=1}(\hat{Y}_{X_1,X_2})}{LZ_{d=1}(Y)}.$$

Thus,

$$LZ_{d=1}(\hat{Y}_{X_1,X_2,X_3}) = LZ_{d=1}(\hat{Y}_{X_1}) + LZ_{d=1}(\hat{Y}_{X_1,X_2}) - LZ_{d=1}(\hat{Y}_{X_1}) + LZ_{d=1}(\hat{Y}_{X_1,X_2,X_3}) - LZ_{d=1}(\hat{Y}_{X_1,X_2})$$

which proves the desired identity.

We finally remark that the previous results hold for non-negative variables. The restriction to non-negative variables is due to the Lorenz Zonoid construction, which in the univariate context corresponds to the Gini measure. As discussed by Raffinetti, Siletti and Vernizzi (2015) the main problem with negative values concerns the violation of the Gini coefficient normalization principle. Indeed, if the negative observed and/or linear estimated values of Y are involved, the corresponding Gini coefficient may achieve values greater than one. This is due to the graphical position of the underlying Lorenz curve partially lying under the x-axis and consequently intersecting the x-axis in a point which defines the transition from cumulative negative values to cumulative non-negative values. Roughly speaking, the Lorenz Zonoid of a real-valued variable is not inscribed into the

unit side square. Moreover, the presence of negative values may lead to failure of the inclusion property. As an example, suppose that $\hat{Y}$ is real-valued and Y is non-negative. In this scenario, the inclusion property would be partially reversed yielding the condition $LZ(Y)_{d=1} \subset LZ(\hat{Y})_{d=1}$ to be fulfilled until the cumulative function of the negative values becomes positive and grater than the cumulative function of the response variable Y values. To overcome this drawback, the Gini coefficient (Lorenz Zonoid) of the real-valued variable has to be adjusted to ensure that its range is bounded between 0 and 1. This adjustment was suggested by Raffinetti et al. (2015) and is based on a new definition of the polarization phenomenon, according to which the total negative variable amount should be assigned to one unit and the remaining total positive variable amount to another unit, while setting the values of the other $n-2$ units equal to 0. In this way, the non-negative variable Lorenz Zonoid and the real-valued variable Lorenz Zonoid become equally scaled.

4 Application

To better understand our proposal, we first consider an illustrative example and then an application to real data.

4.1 Illustrative example

Consider the data in Table 1. This table contains information about the response variable Y and six explanatory variables $(X_1, \ldots, X_6)$, among which X_1 is a nominal variable; X_2, X_3 are quantitative variables and X_4, X_5, X_6 are ordinal variables.

Table 1. A data example with one independent variable Y and six explanatory variables.

Y	23	26	23	28	21	19	19	35	27	11	22	22	26	24	24	26	21	32	17
X_1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
X_2	9	8	7	8	8	8	9	10	9	4	8	9	8	8	7	8	9	10	8
$X3$	10	7	8	8	7	8	8	10	8	7	8	9	8	10	6	8	10	10	9
$X4$	3	2	3	3	3	3	2	3	3	3	3	3	3	3	3	3	3	3	3
$X5$	4	2	6	4	4	6	4	6	6	2	4	4	4	4	2	4	4	4	4
$X6$	4	4	6	4	6	6	6	6	6	2	4	6	4	4	4	4	4	6	4

The data in Table 1 can be used to show how a model selection procedure can be based on Lorenz Zonoids.

First, using the MGC measure we can assess which explanatory variable in the multiple linear regression model is the most important in explaining the (mutual) variability of the response variable. It turns out that X_2 is the most important variable, with $MGC_{(y|x_2)} \simeq 0.599$ meaning that the Lorenz Zonoid of $\hat{Y}_{X_2}$ represents almost the 60% of the Y Lorenz Zonoid. Consequently, the first explanatory variable to be added into the linear regression model is X_2.

Second, among the remaining covariates, the one which gets the largest partial contribution to the variability of the response can be determined computing $PGC_{y,x_?|x_2}$. It turns out that the highest PGC value is obtained for X_5, being $PGC_{y,x_5|x_2} \simeq 0.224$. The inclusion of the covariate X_5 into a model that already has X_2 allows to increase the dilation of $LZ_{d=1}(\hat{Y}_{X_2})$ by 2.4%.

The procedure can then be repeated for all other variables. The results are as follows:

- the third added covariate is X_1, which provides an increase of the dilation of $LZ_{d=1}(\hat{Y}_{X_2,X_5})$

by 15.6% ($PGC_{y,x_1|x_2,x_5} \simeq= 0.156$);

- the fourth added covariate is X_4, which provides an increase of the dilation of $LZ_{d=1}(\hat{Y}_{X_2,X_5,X_1})$ by 14% ($PGC_{y,x_4|x_2,x_5,x_1} \simeq= 0.140$);

- the fifth added covariate is X_3, which provides an increase of the dilation of $LZ_{d=1}(\hat{Y}_{X_2,X_5,X_1,X_4})$ by 9.1% ($PGC_{y,x_3|x_2,x_5,x_1,x_4} \simeq= 0.091$);

- the last added covariate is X_6, which provides an increase of the dilation of $LZ_{d=1}(\hat{Y}_{X_2,X_5,X_1,X_4,X_3})$ by 7.8% ($PGC_{y,x_6|x_2,x_5,x_1,x_4,x_3} \simeq 0.078$).

4.2 Cryptocurrency price data

We consider an application to real data, to illustrate the functioning of our proposed methodology in a professional context. The application concerns the cryptocurrency data collected and illustrated in a recent work of Giudici and Abu-Hashish (2019). We apply our proposal to assess whether the daily bitcoin prices in different crypto exchanges may be affected by the prices of classical financial assets.

The available data contains information on the daily bitcoin prices in eight different crypto exchanges, from 18 May, 2016 to 30 April, 2018. The analysis was carried out including all eight crypto exchanges in Giudici and Abu-Hashish (2019). For the sake of brevity we only report the findings for Coinbase Bitcoin and HitBtc Bitcoin, which represent the response variables of interest. The other exchanges have a similar behavior, due to common underlying bitcoin price.

The explanatory variables which are taken into account are the price of Oil and Gold, that are "classical" financial variables. We first compute the MGC coefficients for both the response variables. Through the MGC coefficients we can detect which covariate provides

the greatest contribution in explaining the bitcoin price mutual variability. Such covariate will be included into the linear regression model. The contribution of the remaining explanatory variable is assessed in terms of the PGC indices. The results from the Lorenz Zonoid measure referred to Coinbase Bitcoin and HitBtc Bitcoin, together with the MGC coefficients, are displayed in Table 2. The table also reports the corresponding values of the coefficient of determination.

Table 2. Lorenz Zonoids, MGC and R^2 measures for Coinbase Bitcoin and HitBtc Bitcoin prices

Target variable	$LZ_{d=1}(\cdot)$	$MGC_{(\cdot\|Gold)}$	$MGC_{(\cdot\|Oil)}$	$R^2_{\cdot,Gold}$	$R^2_{\cdot,Oil}$
Coinbase Bitcoin	0.554	0.398	0.332	0.127	0.086
HitBtc Bitcoin	0.554	0.406	0.341	0.134	0.092

Table 2 shows that both the Coinbase Bitcoin and HitBtc Bitcoin prices present quite the same variability, measured by the corresponding Lorenz Zonoids. We can conclude that both prices do not suffer from strong daily differences. The explanatory variable Gold provides a contribution equal to the 39.8% and 40.6% for the Coinbase Bitcoin and HitBtc Bitcoin variables, respectively. The other covariate, Oil, provides a smaller contribution. Thus, variable Gold is the first variable to be introduced into the model. The results obtained applying the coefficient of determination are similar.

We now consider the effect of introducing an additional variable into a model, comparing our proposed PGC measure with the partial correlation coefficient.

The results of this analysis are displayed in Table 3, for the response variables Coinbase Bitcoin and HitBtc Bitcoin.

From Table 3, note that adding the variable Oil to the model that contains Gold leads

Table 3. PGC and R^2-based indices for Coinbase Bitcoin and HitBtc Bitcoin

Target variable	Additional Covariate	$PGC_{\cdot,Oil\|Gold}$	$r^2_{\cdot,Oil\|Gold}$
Coinbase Bitcoin	Oil	0.236	0.125
HitBtc Bitcoin	Oil	0.248	0.134

to a contribution equal to 23.6% and 24.8% for the Coinbase Bitcoin and HitBtc Bitcoin variables, respectively. Similarly, in terms of the squared partial correlation coefficient, the variable Oil explains the 12.5% and 13.4% of the variance not explained by the model built using only the Gold variable. Similar findings can be obtained for the other crypto exchanges. We can thus conclude that not only the gold price but also the oil price have an important role in the explanation of the bitcoin prices from all exchanges. This conclusion is consistent with what was found by Giudici and Abu-Hashish (2019).

5 Conclusions

In this paper Lorenz Zonoids were introduced as a new model selection tool to assess the contribution associated with the explanatory variables included in a linear model in terms of the explained mutual variability.

Our approach presents similarities and dissimilarities with R^2-based approaches. On the one hand both methods are built on a quantitative response variable and aim to detect the variables which mainly impact the phenomenon of interest. On the other hand our proposal is based on the mutual distance between all observations, rather than deviations from the mean and, therefore, is more robust to outlying observations.

Our proposed Lorenz Zonoid-based model selection approach seems to be a useful model

selection tool, that can be used along with standard tools, to enhance the robustness of the results.

Further research development concerns the extension of the proposed work to a more general class of models that also include non continuous response variables. From an applied viewpoint, it would be interesting to apply the methodology to other related application fields, such as credit scoring (as in Figini and Giudici, 2013 and Calabrese and Giudici, 2015) and operational risk management (as in Fantazzini et al., 2008 and Giudici and Bilotta, 2004).

6 Acknowledgments

We would like to thank the anonymous referees and the editor for their useful comments and suggestions on the paper. The research in the paper was funded by the European Unions Horizon 2020 research and innovation programme under grant agreement No 825215 (Topic: ICT-35-2018 Type of action: CSA). While the paper is the result of a close cooperation between the two Authors, ER wrote Sections 2,3 and 4.1; PG wrote Sections 1,4.2 and 5.

References

BURNHAM , K. P., and ANDERSON, D.R. (2004), "Multimodel Inference: Understanding AIC and BIC in Model Selection", *Sociological Methods & Research*, *33(2)*, 261–304.

CALABRESE, R. and GIUDICI, P. (2015), "Estimating bank default with generalised extreme value models", *Journal of the Operational Research Society*, *66*, 1783–1792.

DALL'AGLIO, M., and SCARSINI, M. (2003), "Zonoids, Linear Dependence, and Size-Biased Distributions on the Simplex", *Advances in Applied Probability*, *35*, 871–884.

DIEBOLD, F.X., and MARIANO, R.S. (1995), "Comparing predictive accuracy", *Journal of Business and Economic Statistics*, *13(3)*, 253–263.

FANTAZZINI, D., DALLA VALLE, L. and GIUDICI, P. (2008), "Copulae and operational risk", *International Journal of Risk Assessment and Management*, *9*, 238-257.

FIGINI, S. and GIUDICI, P. (2015), "Statistical merging of rating models", *Journal of the Operational Research Society*, *62*, 1067–1074.

GIUDICI, P. (2003), *Applied Data Mining: Statistical Methods for Business and Industry* (1st ed.), Hoboken, NJ, USA: Wiley.

GIUDICI, P., and BILOTTA, A. (2004), "Modelling operational losses: a Bayesian approach", *Quality and Reliability Engineering International*, *20*, 407-417.

GIUDICI, P., and RAFFINETTI, E. (2011), "On the Gini measure decomposition", *Statistics & Probability Letters*, *81(1)*, 133–139.

GIUDICI, P., and ABU-HASHISH, I. (2019), "What determines bitcoin exchange prices? A network VAR approach", *Finance Research Letters*, *28*, 309–318.

HAND, D., MANNILLA, H., and SMYTH, P. (2001), *Principles of data mining. Adaptive computation and machine learning series* (1st ed.), Boston: MIT Press.

KOSHEVOY, G. (1995), "Multivariate Lorenz majorization", *Social Choice and Welfare*, *12(1)*, 93–102.

KOSHEVOY, G., and MOSLER, K. (1996), "The Lorenz Zonoids of a Multivariate Distribution", *Journal of the American Statistical Association*, *91(434)*, 873–882.

KOSHEVOY, G., and MOSLER, K. (2007), "Multivariate Lorenz dominance based on Zonoids", *AStA Advances in Statistical Analysis*, *91(1)*, 57–76.

LERMAN, R., and YITZHAKI, S. (1984), "A note on the calculation and interpretation of the Gini index", *Economics Letters*, *15(3-4)*, 363–368.

LORENZ, M.O. (1905), "Methods of Measuring the Concentration of Wealth", *Journal of the American Statistical Association*, *9(70)*, 209–219.

MOSLER, K. (1994), "Majorization in economic disparity measures", *Linear Algebra and its applications*, *119(1)*, 91–114.

MULIERE, P., and PETRONE, S. (1992), "Generalized Lorenz curve and monotone dependence orderings", *Metron L(3-4)*, 19–38.

RAFFINETTI, E., and GIUDICI, P. (2013), "Lorenz Zonoids and Dependence Measures: A Proposal", in Torelli, N., Pesarin, F. and Bar-Hen, A. (eds), *Theoretical and Applied Statistics, Series: Studies in Theoretical and Applied Statistics*, Berlin Heidelberg: Springer.

RAFFINETTI, E., SILETTI, E., and VERNIZZI, A. (2015), "On the Gini coefficient normalization when attributes with negative values are considered", *Statistical Methods & Applications*, *24(3)*, 507–521.

ROUSSEEUW, P.J., and LEROY, A.M. (1987), *Robust regression and outlier detection*, Inc. New York, NY, USA: Wiley.