

Lorenz zonoids and dependence measures: a proposal

Paolo Giudici, Emanuela Raffinetti

Abstract During these last years, the dependence analysis context has assumed a relevant role either in economical and statistical applications: the literature provides a wide set of statistical tools focused in obtaining information about the dependence problem. In this paper the idea consists in focusing the attention on the Lorenz zonoid tool: when considering only the univariate case, the Lorenz zonoid corresponds to the Gini measure.

Our aim is extending the Lorenz zonoids application to the multivariate dimension. In particular we first consider the Lorenz zonoid of a linear regression function characterized by k explanatory variables and then we define the partial contribution due to the introduction of a $(k + 1)$ explanatory variable in terms of dependence measures. The evident effect of a new explanatory variable introduction in the model is translated into the increase of the dilation measure. The final result is characterized by the definition of new dependence measure we called "*partial Lorenz dependence measure*".

Key words: Lorenz and lift zonoid, dilation measure, partial Lorenz dependence measures.

1 Introduction

The dependence research field continues developing thanks to its important contribution to the analysis of specific problems in many application areas: statistical dependence is a type of relation between any two features of units under study.

Paolo Giudici

University of Pavia, Via Strada Nuova 65, e-mail: giudici@unipv.it

Emanuela Raffinetti

University of Pavia, Via Strada Nuova 65, e-mail: emanuela.raffinetti@unipv.it

These units may, for instance, be individuals, or objects, or various aspects of the environment.

The current statistical literature describes different approaches finalized to the study of dependence and concordance problem. Many of the notions of positive dependence are defined by means of some comparison of the joint distribution of two random variables, X and Y , with their distribution under the theoretical assumption that X and Y are independent. Often such a comparison can be extended to general pairs of bivariate distributions with given marginal: this fact led researchers to introduce various notions of positive dependence orderings. The role of these orders consists in comparing the strength of the positive dependence of the two underlying bivariate distributions (see e.g. [1]).

Since the dependence analysis is strictly related to concordance studies, an useful tool focused in getting information about the dependence degree among the involved variables is the so called ranks-based approach (see e.g. [11] and [3]).

Recalling that the existent literature provides a large number of families of bivariate distributions with a natural interpretation of monotone dependence, the intuitive meaning of monotone dependence for a bivariate random vector (X, Y) is that large values of Y stochastically correspond to large values of X (positive dependence) or, in the opposite case, large values of Y correspond to small values of X (negative dependence) (see e.g. [12]).

Different notions of monotone dependence are defined through the conditional distribution of $Y|X = x$ or through the regression function $E(Y|X = x)$: examples can be found in regression dependence (see [9]) and in monotone regression (see e.g. [13]).

In order to compare two bivariate distributions having the same pairs of marginals to determine whether one distribution is more positively dependent than the other, several partial orderings on the class of bivariate distributions with fixed marginals have been introduced. In this field of research a new characterization of monotone dependence in drawing a comparison between the Lorenz curve of the regression function $E(Y|X)$ and the Lorenz curve of Y has been introduced (see e.g. [12]): this notion is appropriate when asking the relation to be invariant under increasing transformation of X but sensible to increasing transformation of Y .

A partial ordering of monotone dependence on the class of nonnegative bivariate random vectors with given marginals has been then defined: the motivation of the proposed partial ordering arises in the study of economic problems (e.g. taxation, see [11]) and in the study of several applications such as discriminant problems and statistical quality control.

2 Dependence measures based on lift and Lorenz zonoids

Let us start introducing the main topics concerning the dependence study: the most important tool used in this context of analysis is recognized in Lorenz and lift zonoids. When considering multivariate data, the Lorenz zonoid represents a spe-

cific characterization of the Lorenz curve. The Lorenz zonoid has been introduced by Koshevoy (see e.g. [4] and [5]) for empirical distributions and by Mosler (see e.g. [10]) for general probability distributions. Our attention will be also focused on the so called lift zonoids application: the lift zonoid of a d -variate measure corresponds to a convex set in $\mathbb{R}^{d+1}$ whose role is directed in analyzing and comparing random vectors. Through the lift zonoid representation one can establish an ordering of random vectors that reflects their variability: in fact, our interest is a comprehensive investigation of the ordering among random vectors that is induced by the inclusion of lift and Lorenz zonoids. Furthermore, starting from the multivariate Lorenz curve generalization one can also define the multivariate Lorenz dominance via inclusion of the Lorenz zonoids.

In Subsection 2.2 we will show that the lift zonoids inclusion allows to define dependence measures in a bivariate context, while in the multivariate case this property will be exploited in order to survey the partial dependence measures recurring to the tool of the Lorenz zonoids and their inclusion when considering a linear regression model (see Section 3).

2.1 Lift and Lorenz zonoids

In many applications it is convenient to employ random variables and general probability distributions rather than empirical distributions only. Already, Max Lorenz discussed his curve for continuous distributions. The Lorenz curve of a random variable X having finite expectation $\mu > 0$ is the graph of the function (see e.g. [6])

$$L_X(t) = \frac{1}{\mu} \int_0^t F_X^{-1}(s) ds, \quad 0 \leq t \leq 1, \quad (1)$$

where F_X^{-1} is the quantile function of X , $F_X^{-1}(t) = \min\{x : F(x) \geq t\}$, $0 < t \leq 1$.

The lift zonoid of a general d -variate random vector is defined as follows (see e.g. [7]). Consider the set $\mathcal{X}^d$ of random vectors in $\mathbb{R}^d$ that have finite expectation, the subset $\mathcal{X}^{d+} \subset \mathcal{X}^d$ of those vectors that have positive (in each component) expectation, and the subset $\mathcal{X}_+^{d+} \subset \mathcal{X}^{d+}$ of those that have, in addition, support in $\mathbb{R}_+^d$. Define, for $\mathbf{X} \in \mathcal{X}^d$, the lift zonoid

$$\begin{aligned} \hat{Z}(\mathbf{X}) &= \{E[h(\mathbf{X}), h(\mathbf{X})\mathbf{X}] : h : \mathbb{R}^d \rightarrow [0, 1] \text{ measurable}\} \\ &= \{\mathbf{z} \in \mathbb{R}^{d+1} : \mathbf{z} = (z_0, z_1, \dots, z_d) = \zeta(h), h : \mathbb{R}^d \rightarrow [0, 1] \text{ measurable}\}, \end{aligned} \quad (2)$$

where

$$\zeta(h) = \left(\int_{\mathbb{R}^d} h(\mathbf{x}) dF(\mathbf{x}), \int_{\mathbb{R}^d} \mathbf{x} h(\mathbf{x}) dF(\mathbf{x}) \right). \quad (3)$$

The lift zonoid is a convex compact set in $\mathbb{R}^{d+1}$ that includes the expectations of all d -variate random vectors $(g(\mathbf{X}), g(\mathbf{X})\mathbf{X})$.

The lift zonoid has many attractive properties, which makes it useful for a broad range of applications. One property is uniqueness: like the generalized Lorenz curve in dimension one, the lift zonoid contains full information about the underlying distribution, that is, given the lift zonoid, the data can be completely regained. Another important property is that the lift zonoid is additive in distribution. But, in this context, the most important property which will be subsequently recalled, is the *marginalization property*.

Property 1. For any $J \subset \{1, 2, \dots, d\}$ consider the projection $pr_J : \mathbf{z} \rightarrow (z_0, \mathbf{z}_J), \mathbf{z} \in \mathbb{R}^{d+1}$, where $\mathbf{z}_J$ is the subvector containing components $z_j, j \in J$. Then $pr_J(\hat{Z}(\mathbf{X})) = \hat{Z}(\mathbf{X}_J)$.

The general definition of the Lorenz zonoid of a random vector and its properties are straightforward. The Lorenz zonoid of some $\mathbf{X} \in \mathcal{X}_+^d$ is the lift zonoid of the relative vector, i.e. the vector componentwise divided by its expectation. With

$$\tilde{\mathbf{X}} = \left(\frac{X_1}{E[X_1]}, \dots, \frac{X_d}{E[X_d]} \right)$$

define

$$\begin{aligned} LZ(\mathbf{X}) = \hat{Z}(\tilde{\mathbf{X}}) &= \{E[h(\tilde{\mathbf{X}}), h(\tilde{\mathbf{X}})\tilde{\mathbf{X}}] : h : \mathbb{R}_+^d \rightarrow [0, 1] \text{ measurable}\} \\ &= \{\mathbf{z} \in \mathbb{R}^{d+1} : \mathbf{z} = (z_0, z_1, \dots, z_d) = \zeta(h), h : \mathbb{R}_+^d \rightarrow [0, 1] \text{ measurable}\}, \end{aligned} \quad (4)$$

where

$$\zeta(h) = \left(\int_{\mathbb{R}_+^d} h(\mathbf{x}) dF(\mathbf{x}), \int_{\mathbb{R}_+^d} \psi_F(\mathbf{x}) h(\mathbf{x}) dF(\mathbf{x}) \right), \quad (5)$$

and $\psi_F(\mathbf{x}) = \tilde{\mathbf{x}} = (\tilde{x}_1, \dots, \tilde{x}_d), \tilde{x}_j = \frac{x_j}{\mu_j}$, for $j = 1, \dots, d$. The Lorenz zonoid of a random vector $\mathbf{X} \in \mathcal{X}_+^d$ is a convex compact set in $\mathbb{R}^{d+1}$.

2.2 Lift zonoids approach: the bivariate case

Let us introduce what gives our construction in the dimension two (see e.g. [8]): namely, let (X, Y) be a bivariate random vector with distribution $F(x, y)$. The lift zonoid $\hat{Z}(F)$ is a convex compact in $\mathbb{R}^3$ by definition: the projection $pr_J(\hat{Z}(F))$ is the lift zonoid $\hat{Z}(F^1)$, which is the area between the generalized Lorenz curve of variable X and the generalized dual Lorenz curve of variable X , denoted by $L'_X(t) = E(X) - L_X(1 - t)$, where $L_X(t)$ represents the generalized Lorenz curve which corresponds to the graph of points $(\int_{-\infty}^a dF^1(x), \int_{-\infty}^a x dF^1(x))$, with $a \in (-\infty, +\infty)$, in $\mathbb{R}^2$.

Let us calculate the inverse image $pr_1^{-1}(L_X(t))$: $pr_1^{-1}(L_X(t)) = (\int_{-\infty}^a dF^1(x), \int_{(x,y):x \leq a} xF(dx, dy), \int_{(x,y):x \leq a} yF(dx, dy))$. The projection of this point under pr_2 is equal to $(\int_{-\infty}^a dF^1(x), \int_{(x,y):x \leq a} yF(dx, dy))$, so that the latter point obviously equals the collection of points $(\int_{-\infty}^a dF^1(x), \int_{-\infty}^a E(Y|X=x)F^1(dx))$: in fact, $F(dx, dy) = F(dy|dx)F^1(dx)$ and so substituting it in $\int_{(x,y):x \leq a} yF(dx, dy)$, we get $\int_{-\infty}^a yF(dy|dx)F^1(dx) = \int_{-\infty}^a E(Y|X=x)F^1(dx)$. All these points define the generalized Lorenz curve, $L_{E(Y|X)}$, of $E(Y|X)$ when a ranges over $(-\infty, +\infty)$.

Similarly, the inverse image of the generalized dual Lorenz curve is the collection of the form $(\int_a^{+\infty} F^1(dx), \int_a^{+\infty} E(Y|X=x)F^1(dx))$, when $a \in (-\infty, +\infty)$, denoted by $L'_{E(Y|X)}$.

Generally speaking, those curves have not to be concave or convex: consider the convex hull of the collection of points that characterizes $L_{E(Y|X)}$ and $L'_{E(Y|X)}$ and denote it by $\hat{Z}(E(Y|X))$. This set is centrally symmetric, therefore, since we are interested to the dimension 2 context analysis, there exists a distribution, indicated by $\mu_{E(Y|X)}$, on the real line, such that $\hat{Z}(\mu_{E(Y|X)}) = \hat{Z}(E(Y|X))$. We call the measure $\mu_{E(Y|X)}$ the *dependency measure* of Y on X .

3 Lorenz zonoids-based approach: the multivariate case

Up to now the dependence study has been surveyed only with reference to bivariate random vectors (see e.g [8]): the question is how extending the dependence analysis when considering a linear regression function characterized by more than one explanatory variable.

The starting point is characterized by defining a single partial dependence measure whose role consists in expressing the "share" of the Y Lorenz zonoid "explained" by each single considered explanatory variable. Of our interest are the following computations. Let us define the response variable Y Lorenz zonoid intended as the area between the Lorenz curve and its dual and denote it by $A(Y)$: when $d = 1$, this area corresponds to the Gini measure, implying in this second study setting, the procedure will be simplified by bringing back all the analysis to the univariate context.

Assuming the application of a linear regression function and once computing the estimated Y variable values only considering the explanatory variable X_1 , one can build the Lorenz zonoid of $\hat{Y}_{X_1} = E(Y|X_1)$ and denote it by $A(\hat{Y}_{X_1})$. Then, the subsequent step, recalling that for the sake of simplicity we are going to operate only with a trivariate random vector (Y, X_1, X_2) , concerns the computation of the Lorenz zonoid of the estimated Y variable values according to the second explanatory variable X_2 , obtaining $A(\hat{Y}_{X_2})$. Finally, exploiting the multiple linear regression tool, one can get the estimated Y variable values $\hat{Y}_{X_1, X_2} = E(Y|X_1, X_2)$, considering both the two explanatory variables X_1 and X_2 : the area that ranges between the Lorenz curve of $\hat{Y}_{X_1, X_2}$ and its dual is denoted by $A(\hat{Y}_{X_1, X_2})$.

These areas are particularly useful in order to obtain partial dependence mea-

asures: intuitively a partial measure representative the share of the Y Lorenz zonoid "explained" by the only variable X_1 , is derived through the following ratio

$$G_1 = \frac{A(\hat{Y}_{X_1})}{A(Y)}, \quad (6)$$

and, in the same manner, a partial measure able to describe the share of the Y variable Lorenz zonoid "explained" by the only variable X_2 , is derived by the ratio

$$G_2 = \frac{A(\hat{Y}_{X_2})}{A(Y)}. \quad (7)$$

According to the purpose of getting partial dependence measures expressed in terms of Lorenz zonoids, since we are using the linear regression model, the idea consists in determining the partial contribution of each explanatory variable to the Y variability explanation.

Suppose, in general, to consider k explanatory variables: the aim is in establishing the effect, in terms of dependence measures, connected to the addition of a new $(k+1)$ variable in the model. Intuitively, the addition of a new variable in the model can provide the effect of an enlargement of the $E(Y|\cdot)$ Lorenz zonoid. The Lorenz zonoid of the Y estimated values on the basis of the k explanatory variables and denoted by $LZ(E(Y|X_1, \dots, X_k))$, corresponds to the dilation measure of the Y response variable Lorenz zonoid $LZ(Y)$. In other words, the introduction of an additional explanatory variable in the linear regression model translates into an increment of the "explained" model variance: this result is confirmed by the Lorenz zonoids inclusion. The problem, then, is how quantifying the partial contribution due to the new added variable in terms of Lorenz zonoids: the idea consists in retracing some existing results concerning the general regression model (see e.g. [2]).

In this direction the partial determination coefficient square formula

$$r_{Y, X_{k+1}|X_1, \dots, X_k}^2 = \frac{\text{Var}(\hat{Y}_{X_{k+1}}) - \text{Var}(\hat{Y}_{X_k})}{\text{Var}(Y) - \text{Var}(\hat{Y}_{X_k})} \quad (8)$$

is very useful in order to establish the additional contribution in terms of the explained variance and the variability not explained by $\hat{Y}_{X_k}$.

Coming back to our original analysis, it's trivial to show the existence of a relationship between the partial correlation coefficient and the dependence measure based on the Lorenz zonoids.

The additional contribution related to the $(k+1)$ variable introduction can be measured by the difference between the Lorenz zonoid of $\hat{Y}_{X_{k+1}}$ and $\hat{Y}_{X_k}$, that is

$$A(\hat{Y}_{X_{k+1}}) - A(\hat{Y}_{X_k}), \quad (9)$$

thus, a relative measure able to stress the additional contribution of the X_{k+1} variable is provided by

$$RG = \frac{A(\hat{Y}_{X_{k+1}}) - A(\hat{Y}_{X_k})}{A(Y) - A(\hat{Y}_{X_k})}. \quad (10)$$

This relative measure will be called *partial Lorenz dependence measure*. Let us provide a proof of (10) recurring to the multiple regression correlation coefficient formula and in particular when two explanatory variables are involved.

Proof. In the case that only two explanatory variables are considered, the multiple regression correlation coefficient assumes the following expression

$$R^2 = r_{Y,X_1}^2 + r_{Y,X_2|X_1}^2 (1 - R_{Y,X_1}^2). \quad (11)$$

In a linear regression model the multiple determination coefficient R^2 defines the linear dependence among all the involved variables: in terms of Lorenz zonoids the total dependence measure corresponds to $LZ(E(Y|X_1, X_2))$ which is equivalent to the area $A(\hat{Y}_{X_1, X_2})$ between the Lorenz curve of $\hat{Y}_{X_1, X_2} = E(Y|X_1, X_2)$ and its dual. In terms of Lorenz zonoids one gets

$$LZ(E(Y|X_1, X_2)) = LZ(E(Y|X_1)) + LZ(E(Y|(X_2|X_1)))(LZ(Y) - LZ(E(Y|X_1))),$$

where $LZ(E(Y|X_1)) = A(\hat{Y}_{X_1})$, $LZ(E(Y|(X_2|X_1))) = RG$ and $LZ(Y) = A(Y)$. Now, substituting to RG its expression in (10), one gets

$$\begin{aligned} LZ(E(Y|X_1, X_2)) &= A(\hat{Y}_{X_1}) + RG(A(Y) - A(\hat{Y}_{X_1})) \\ &= A(\hat{Y}_{X_1}) + \frac{A(\hat{Y}_{X_1, X_2}) - A(\hat{Y}_{X_1})}{A(Y) - A(\hat{Y}_{X_1})} (A(Y) - A(\hat{Y}_{X_1})) \\ &= A(\hat{Y}_{X_1}) + A(\hat{Y}_{X_1, X_2}) - A(\hat{Y}_{X_1}) \\ &= A(\hat{Y}_{X_1, X_2}) \quad \square \end{aligned}$$

The last topic, to be described, is the statistical interpretation to assign to partial Lorenz dependence measures: we will consider a very simple example that combines both the main topics linked to the linear regression model and the Lorenz zonoids theory. Suppose to consider these data:

Table 1 Data

Y	X ₁	X ₂
350	84	45
202	73	19
404	64	53
263	68	31
451	76	58

Using the multiple linear regression model and the Lorenz zonoid tool, we get $A(Y) \cong 0.153$, $A(\hat{Y}_{X_1}) \cong 0.043$, $A(\hat{Y}_{X_2}) \cong 0.023$ and $A(\hat{Y}_{X_1, X_2}) \cong 0.054$. Through

(6) and (7), single partial dependence measures can be derived so that $G_1 \cong 0.279$ and $G_2 \cong 0.152$: the relative interpretation is very simple. The Lorenz zonoid of $E(Y|X_1) = G_1$ represents the 27.9% of the Lorenz zonoid of Y and the Lorenz zonoid of $E(Y|X_2) = G_2$ represents the 15.2% of the Lorenz zonoid of Y . The relative measure describing the additional contribution of the variable X_2 to the Lorenz zonoid of Y is $RG_{X_1, X_2} \cong 0.105$ meaning that the introduction of the explanatory variable X_2 allows to increase the dilation of the Lorenz zonoid of $E(Y|X_1)$ in measure equivalent to 10.5%. Finally, the relative measure describing the additional contribution of the variable X_1 to the Lorenz zonoid of Y is $RG_{X_2, X_1} \cong 0.238$ meaning that the introduction of the explanatory variable X_1 in the model allows to increase the dilation of the Lorenz zonoid of $E(Y|X_2)$ in measure equivalent to 23.8%. Thus, we can conclude that an increase of the dilation measure implies a reduction of the unexplained linear regression model variance.

Acknowledgements The authors acknowledge financial support from the European grant EU-IP MUSING (contract number 027097). The paper is the result of the close collaboration among the authors, however, it has been written by Emanuela Raffinetti with the supervision of Prof. Paolo Giudici.

References

1. Colangelo, A., Scarsini, M., Shaked, M.: Some Positive Dependence Stochastic Orders. *Studi statistici* No.84, Istituto di Metodi quantitativi, Università Luigi Bocconi (2004).
2. Giudici, P.: *Data Mining. Metodi informatici, statistici e applicazioni*. McGraw Hill (2005)
3. Giudici, P., Raffinetti, E.: Multivariate Ranks-based concordance indexes. *Convegno Società Italiana di Statistica*, 23-25 Settembre, Pescara (2009)
4. Koshevoy, G.: Multivariate inequality indices and orderings on a product of symmetric groups. *Economica i Matematicheskie Methody*, 29 (1993)
5. Koshevoy, G.: Multivariate Lorenz majorization. *Social Choice and Welfare*, 12 (1995)
6. Koshevoy, G., Mosler, K.: The Lorenz Zonoids of a Multivariate Distribution. *Journal of the American Statistical Association*, 91, No. 434, Theory and Methods (1996)
7. Koshevoy, G., Mosler, K.: Multivariate Lorenz dominance based on zonoids. *AStA Advance in Statistical Analysis*, 91, No. 1, 57–76 (2007)
8. Koshevoy, G., Muliere, P.: Dependence orderings. Unpublished draft (2002)
9. Lehmann, E.L.: Some concepts of dependence. *Ann. Math. Statist.*, 37, (1966)
10. Mosler, K.: Majorization in economic disparity measures. *Linear Algebra and its applications*, 220 (1994)
11. Muliere, P.: Alcune osservazioni sull'equità orizzontale di una tassazione. *Scritti in onore di Francesco Brambilla*. Ed. by Bocconi Comunicazione 2, (Milano, 1986)
12. Muliere, P., Petrone, S.: Generalized Lorenz curve and monotone dependence orderings. *Metron Vol L*, No. 3–4 (1992)
13. Shea, G.: Monotone regression and covariance structure. *Annals of Statistics*, Vo. 7, (1979)