DISCUSSION PAPERS IN STATISTICS AND ECONOMETRICS

SEMINAR OF ECONOMIC AND SOCIAL STATISTICS
UNIVERSITY OF COLOGNE

No. 7/95

Multivariate Gini Indices

by

G.A. Koshevoy and K. Mosler*

June 1995

Abstract

The Gini index and the Gini mean difference of a univariate distribution are extended to measure the disparity of a general d-variate distribution. We propose and investigate two approaches, one based on the distance of the distribution from itself, the other on the volume of a convex set in $(d+1)$-space, named the lift zonoid of the distribution. When $d = 1$, this volume equals the area between the usual Lorenz curve and the line of zero disparity, up to a scale factor. We get two definitions of the multivariate Gini index, which are different (when $d > 1$) but connected through the notion of the lift zonoid. Both notions inherit properties of the univariate Gini index, in particular, they are vector scale invariant, continuous, bounded by 0 and 1, and the bounds are sharp. They vanish if and only if the distribution is concentrated at one point. The indices have a *ceteris paribus* property and are consistent with multivariate extensions of the Lorenz order. Illustrations with data conclude the paper.

Key words: Dilation; Disparity measurement; Gini mean difference; Lift zonoid; Lorenz order.

*Seminar für Wirtschafts– und Sozialstatistik, Universität zu Köln, Meister-Ekkehart-Str. 9/II, D-50923 Köln; Tel: +49/221/470 4266, e–mail: mosler@wiso.uni-koeln.de

1 Introduction

To measure the disparity of a probability distribution, the Gini mean difference and its scale invariant version, the Gini index, are most widely used. The Gini index is closely connected to the Lorenz curve; it amounts to twice the area between the Lorenz curve and the diagonal of the unit square. By this, the Gini index is consistent with the Lorenz order: It increases from one distribution to another if the first Lorenz curve lies above the second.

In this paper we investigate extensions of the Gini mean difference and the Gini index to measure the disparity of a population with respect to several attributes $s = 1, \ldots, d$. The Gini mean difference of a univariate distribution F is defined as the expected distance between two independent random variables which both follow the law F. Our first notion will be an immediate extension of this (Section 4). For a d-variate empirical distribution F_A, with data matrix $A = [a_{is}]$, it reads

$$M_D(F_A) = \frac{1}{2n^2 d} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\sum_{s=1}^{d} (a_{is} - a_{js})^2 \right)^{\frac{1}{2}}. \tag{1}$$

We call M_D the distance–Gini mean difference. Our second notion, M_V, will be based on the volume of the lift zonoid and named the *volume–Gini mean difference* (Section 5). The lift zonoid of a d-variate distribution is a convex set in $\mathbb{R}^{d+1}$ which extends the generalized Lorenz curve; see Section 3 below. For F_A we will get

$$M_V(F_A) = \frac{1}{2^d - 1} \sum_{s=1}^{d} \frac{1}{n^{s+1}} \sum_{1 \le i_1 < \ldots < i_{s+1} \le n} \sum_{1 \le r_1 < \ldots < r_s \le d} |\det(\mathbf{1}, A^{r_1,\ldots,r_s}_{i_1,\ldots,i_{s+1}})|, \tag{2}$$

where $\mathbf{1}$ is a column of ones, and $A^{r_1,\ldots,r_s}_{i_1,\ldots,i_{s+1}}$ is the matrix obtained from the rows $i_1, \ldots, i_{s+1}$ and the columns $r_1, \ldots, r_s$ of the data matrix.

For univariate data, the Gini index equals the Gini mean difference of the relative data, which are the original data 'scaled down' by their mean. Thus, for a d-variate distribution we will define the distance-Gini index and the volume-Gini index by

$$R_D(F_A) = M_D(\widetilde{F}_A) \quad \text{and} \quad R_V(F_A) = M_V(\widetilde{F}_A), \tag{3}$$

where F_A is componentwise scaled down to $\widetilde{F}_A$ by its mean vector; see Section 3.

Every d-variate Gini index should have at least the following properties: be equal to the usual Gini index in case $d = 1$, vary between 0 and 1, be scale invariant, and increase with a proper multivariate extension of the Lorenz order. This and

more will be shown for our two notions. Also they will be investigated for general d-variate probability distributions.

For univariate distributions, the Gini mean difference increases with the dilation order, and the Gini index increases with the Lorenz order, which we call relative dilation because it amounts to dilation of the relative distributions. Of course, dilation implies relative dilation.

We will consider several extensions of dilation to the multivariate case. The first is classical d-variate dilation, which means that one distribution equals the other one plus 'noise'. The second, directional dilation, has the following property. G is a directional dilation of F if and only if the lift zonoid of G includes that of F. Further, absolute and relative versions of these dilations are considered in Section 3. We will show in Section 6 that both notions of the Gini mean difference are increasing with absolute dilation and directional absolute dilation. Similarly, both our Gini indices increase with relative dilation and directional relative dilation.

Although M_D and R_D are obvious extensions of the univariate notions, most of their properties have not been explored so far. In particular we prove in Section 4 that R_D varies between 0 and 1, and that the bounds are sharp. We also establish a connection between $M_D(F)$ and the lift zonoid of F: $M_D(F)$ is proportionate to the average area of certain two–dimensional projections of the lift zonoid (Remark 4.1).

There are several attempts in the literature to define a multivariate Gini mean difference. Wilks (1960) proposes the volume of a convex body associated with F. Oja (1983) shows that the Wilks index is the expected volume of a simplex generated by $d+1$ random vertices which are independent and identically distributed according to F; see also Giovagnoli and Wynn (1995). In our framework, the Wilks index amounts to $d+1$ times the volume of the lift zonoid (Theorem 5.1). Torgersen (1991) uses, as a multivariate Gini mean difference, the volume of the zonoid of the distribution, which is the projection of its lift zonoid on the last d coordinates. For a one-point distribution, both the Wilks–Oja and the Torgersen indices vanish. But also for many other distributions they are zero, which appears to be unsatisfactory. Our notion $M_V(F)$ avoids this drawback; it vanishes if and only if F is a one-point distribution. In addition, we provide the correct scaling factor which makes R_V vary between 0 and 1. $M_V(F)$ is an average of projections of the lift zonoid on coordinate planes (Remark 5.1).

Another multivariate Gini index, associated with a concentration surface, has been introduced by Taguchi (1981). For the relations between Taguchi's concentration surface and the lift zonoid, see Koshevoy and Mosler (1995a).

Overview: Some properties of the usual univariate Gini index will be surveyed in Section 2. Section 3 presents the definitions of six multivariate dilation orderings and

of the lift zonoid and the Lorenz zonoid of a d-variate distribution. Section 4 is about the multivariate distance–Gini index and its properties. The multivariate volume–Gini index is introduced and analyzed in Section 5. In Section 6 we demonstrate that our Gini indices are increasing with multivariate dilations. Section 7 concludes the paper with a numerical illustration.

Notation: $\mathbb{R}^k$ ($\mathbb{R}^k_+$) is the k–dimensional Euclidean space of row vectors (nonnegative row vectors). In $\mathbb{R}^k$, x^T is the transpose of a vector x, $\leq$ the usual componentwise ordering, and S^{k-1} the unit sphere. $\mathbf{0}$ stands for the origin, and $\overline{x,y}$ for the segment between x and y in $\mathbb{R}^k$. $[a_1,\ldots,a_l]$ denotes the $l \times k$ matrix with rows $a_1,\ldots,a_l \in \mathbb{R}^k$. For D and E in $\mathbb{R}^k$, $D+E = \{u : u = x+y, x \in D, y \in E\}$ is the Minkowski sum, and $\mathrm{V}_k(D)$ is the k-dimensional volume of D.

2 The univariate Gini index

We will shortly survey the Gini mean difference and the Gini index of a univariate distribution. Let $F : \mathbb{R} \to [0,1]$ be a given probability distribution function on $\mathbb{R}$ which has a finite expectation $\mu(F) = \int_{-\infty}^{\infty} x dF(x) > 0$.

Definition 2.1 (Gini mean difference, Gini index)

$$M(F) = \frac{1}{2}\int_{\mathbb{R}}\int_{\mathbb{R}} |x-y| dF(x) dF(y). \tag{4}$$

is the Gini mean difference *of* F. $R(F) = M(F)/|\mu(F)|$ *is the* Gini index *of* F.

$M(F)$ is the mean Euclidean distance between two independent random variables divided by two, where both random variables are distributed with F, and $R(F)$ is the mean Euclidean distance divided by twice the expectation of F. The definition and, as we will see in Section 4, the following results hold also for distributions with $\mu(F) < 0$.

Proposition 2.1 *Let* $F^{-1}(s) = \inf\{x : F(x) \geq s\}, s \in]0,1]$, *denote the inverse distribution function of* F, *and* $L_F(t) = \mu(F)^{-1}\int_0^t F^{-1}(s)ds$, $t \in [0,1]$. *Then, if* $F(0) = 0$,
(i) $M(F)$ *equals the area between the graphs of the two functions* $t \mapsto |\mu(F)|\, L_F(t)$ *and* $t \mapsto |\mu(F)|\,(1 - L_F(1-t))$, $t \in [0,1]$.
(ii) $R(F)$ *equals the area between the graphs of the two functions* $t \mapsto L_F(t)$ *and* $t \mapsto 1 - L_F(1-t)$, $t \in [0,1]$.

Proof. $t \mapsto L_F(t)$ is the *Lorenz function*, and its graph is the Lorenz curve of F. $t \mapsto |\mu(F)|\, L_F(t)$ is the *generalized Lorenz function*. It is wellknown that $R(F)$ amounts to twice the area between the Lorenz curve and the main diagonal of the unit square. The area between the main diagonal and the graph of $t \mapsto (1-L_F(1-t))$ is congruent to this first area. Hence (ii). Part (i) follows immediately since $M(F) = |\mu(F)|\, R(F)$. □

The special case of an empirical distribution is particularly important. Let F_a denote the distribution function which gives equal weight to n given points a_i in $\mathbb{R}$, $a_1 \leq \ldots \leq a_n$, $a = (a_1, \ldots, a_n)$, and let $\overline{a} = n^{-1}(a_1 + \ldots + a_n)$. Then the Lorenz curve of F_a is the linear interpolation of the points $(k/n, a_1/\overline{a} + \ldots + a_k/\overline{a})$, $k = 1, \ldots, n$, in two-space.

$$M(a_1, \ldots, a_n) = M(F_a) = \frac{1}{2n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} |a_i - a_j| \tag{5}$$

is the *Gini mean difference* of the *sample* $a = (a_1, \ldots, a_n)$, and

$$R(a_1, \ldots, a_n) = R(F_a) = \frac{1}{\overline{a}} M(a_1, \ldots, a_n) \tag{6}$$

is the *Gini index* of a, provided the sample mean is not zero. The Gini index of a equals the Gini mean difference of the 'scaled down' sample $\widetilde{a} = (a_1/\overline{a}, \ldots, a_n/\overline{a})$,

$$R(a_1, \ldots, a_n) = \frac{1}{2n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \left| \frac{a_i}{\overline{a}} - \frac{a_j}{\overline{a}} \right| . \tag{7}$$

The Gini index and the Gini mean difference have interesting properties which we will extend to our multivariate notions. Here we state them for empirical distributions. They hold as well for general univariate distributions.

Proposition 2.2 (i) *Let* $(a_1, \ldots, a_n) \in \mathbb{R}^n_+$ *with* $\sum a_i > 0$. *Then*

$$0 = R(\overline{a}, \ldots, \overline{a}) \leq R(a_1, \ldots, a_n) \leq R(0, \ldots, 0, \sum_{i=1}^{n} a_i) = 1 - \frac{1}{n} < 1,$$

$$\begin{aligned} R(\beta a_1, \ldots, \beta a_n) &= R(a_1, \ldots, a_n) \quad \textit{for every } \beta > 0, \\ R(a_1 + \lambda, \ldots, a_n + \lambda) &= \frac{\overline{a}}{\overline{a} + \lambda} R(a_1, \ldots, a_n) \quad \textit{for every } \lambda > 0. \end{aligned} \tag{8}$$

(ii) *R is strictly increasing with the Lorenz order, i.e.,*

$R(a_1, \ldots, a_n) > R(b_1, \ldots, b_n)$ if $L_{F_a}(t) \leq L_{F_b}(t)$ for all t and $<$ for some t.

(iii) *R is a continuous function $\mathbb{R}^n \to \mathbb{R}$.*

Proposition 2.3 (i) *Let* $(a_1, \ldots, a_n) \in \mathbb{R}^n_+$ *with* $\sum a_i > 0$. *Then*

$$0 = M(\overline{a}, \ldots, \overline{a}) \leq M(a_1, \ldots, a_n) \leq M(0, \ldots, 0, \sum_{i=1}^{n} a_i) = \overline{a}(1 - \frac{1}{n}) < \overline{a}$$

$$M(\beta a_1, \ldots, \beta a_n) = \beta M(a_1, \ldots, a_n) \text{ for every } \beta > 0$$

$$M(a_1 + \lambda, \ldots, a_n + \lambda) = M(a_1, \ldots, a_n) \text{ for every } \lambda \in \mathbb{R}.$$

(ii) M *is strictly increasing with the Lorenz order.*

(iii) M *is a continuous function* $\mathbb{R}^n \to \mathbb{R}$.

These and other properties have been investigated by many authors. For surveys and references, see Nygård and Sandström (1981) and Giorgi (1990, 1992).

3 Multivariate dilations and the lift zonoid

Let $\mathcal{F}^d$ ($\mathcal{F}^d_0$) be the class of probability distribution functions $\mathbb{R}^d \to \mathbb{R}$ which have a finite (finite and non-zero) expectation vector, and let $\mathcal{F}^d_+ \subset \mathcal{F}^d_0$ be the subclass of probability distributions on the nonnegative orthant $\mathbb{R}^d_+$. Given $F \in \mathcal{F}^d$, let $\mu(F) = \int_{\mathbb{R}^d} x dF(x) = (\mu_1, \ldots, \mu_d)$. For every $F \in \mathcal{F}^d$ and $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$, define $F_{\cdot\beta}(x_1, \ldots, x_d) = F(x_1\beta_1, \ldots, x_d\beta_d)$, and $F_{+\beta}(x_1, \ldots, x_d) = F(x_1 + \beta_1, \ldots, x_d + \beta_d)$.

For $F \in \mathcal{F}^d_0$, $\widetilde{F} = F_{\cdot\mu(F)}$ is called the *relative distribution function,* namely, if F is the distribution function of a random vector $X = (X_1, \ldots, X_d)$, then $\widetilde{F}$ is the distribution of

$$\widetilde{X} = \left(\frac{X_1}{|\mu_1|}, \ldots, \frac{X_d}{|\mu_d|}\right).$$

In the sequel, when using $\widetilde{F}$, we tacitly assume that $F \in \mathcal{F}^d_0$.

Given F and G in $\mathcal{F}^d$, let X and Y be two random vectors from the same probability space which are distributed according to F and G, respectively. G is a *dilation* of F, $F \preceq G$, if there exists a random vector Z such that $E(Z \mid X) = 0$ and Y has the same distribution as $X + Z$. The random variable Z may be interpreted as 'noise', so that Y is distributed like X plus some noise.

We call G an *absolute dilation* of F, $F \preceq_a G$, if, $G_{-\mu(G)}$ is a dilation of $F_{-\mu(F)}$. Given F and G in $\mathcal{F}^d_0$, G is a *relative dilation* of F, $F \preceq_r G$, if, $\widetilde{G}$ is a dilation of $\widetilde{F}$. For $F \in \mathcal{F}^d$ and $p = (p_1, \ldots, p_d) \in \mathbb{R}^d$, we denote

$$F(t, p) = \int_{\{x \in \mathbb{R}^d : xp^T \leq t\}} dF(x), \quad t \in \mathbb{R},$$

$$\widetilde{F}(t,p) = \int_{\{x \in \mathbb{R}^d : xp^T \le t\}} d\widetilde{F}(x), \quad t \in \mathbb{R}.$$

If F is the distribution function of the random vector X in $\mathbb{R}^d$, then $F(\cdot,p)$ is the distribution function of the random variable $p_1X_1 + \ldots + p_dX_d$ in $\mathbb{R}$; similarly $\widetilde{F}(\cdot,p)$ is the distribution function of $p_1X_1/|\mu_1| + \ldots + p_dX_d/|\mu_d|$.

G is a *directional dilation* of F, $F \preceq_{dir} G$, if, for every $p \in S^{d-1}$, $G(\cdot,p)$ is a dilation of $F(\cdot,p)$. We will say that G is a *directional relative dilation* of F, $F \preceq_{dirr} G$, if, for every $p \in S^{d-1}$, $\widetilde{G}(\cdot,p)$ is a dilation of $\widetilde{F}(\cdot,p)$. Similarly, G is named a *directional absolute dilation* of F, $F \preceq_{dira} G$, if, for every $p \in S^{d-1}$, $G(\cdot,p)$ is an absolute dilation of $F(\cdot,p)$.

All these dilations are partial orders (reflexive, transitive and antisymmetric) on $\mathcal{F}^d$, and related by the following implications.

$$\begin{array}{ccc} F \preceq G & \Longrightarrow & F \preceq_{dir} G \\ \Downarrow & & \Downarrow \\ F \preceq_r G & \Longrightarrow & F \preceq_{dirr} G \end{array}$$

$$\begin{array}{ccc} F \preceq G & \Longrightarrow & F \preceq_{dir} G \\ \Downarrow & & \Downarrow \\ F \preceq_a G & \Longrightarrow & F \preceq_{dira} G \end{array}$$

However, in general, no reverse implication holds. For proofs, see Section 6 below. Next we define a multivariate generalization of the Lorenz curve and the generalized Lorenz curve.

Definition 3.1 (Koshevoy and Mosler (1995a,b)) *Let $F \in \mathcal{F}^d$. For a measurable function $h : \mathbb{R}^d_+ \to [0,1]$, consider the vector $(z_0(F,h), z(F,h)) \in \mathbb{R}^{d+1}$, where*

$$z_0(F,h) = \int_{\mathbb{R}^d} h(x)dF(x), \quad z(F,h) = \int_{\mathbb{R}^d} h(x)xdF(x).$$

The set

$$\widehat{Z}(F) \equiv \left\{ (z_0(F,h), z(F,h)) \;\; : \;\; h : \mathbb{R}^d_+ \to [0,1] \;\; \textit{measurable} \right\}$$

is called the lift-zonoid *of F. $LZ(F) = \widehat{Z}(\widetilde{F})$ is called* the Lorenz zonoid *of F.*

The lift zonoid is a multivariate generalization of the generalized Lorenz curve, and the Lorenz zonoid is one of the Lorenz curve. The following theorem establishes the relation between the lift-zonoid and directional dilation.

Theorem 3.1 (Koshevoy and Mosler (1995a,b)) *For $F, G \in \mathcal{F}^d_+$,*
(i) *$F \preceq_{dir} G$ if and only if $\widehat{Z}(F) \subset \widehat{Z}(G)$,*
(ii) *$F \preceq_{dirr} G$ if and only if $LZ(F) \subset LZ(G)$,*
(iii) *$F \preceq_{dira} G$ if and only if $\widehat{Z}(F_{-\mu(F)}) \subset \widehat{Z}(G_{-\mu(G)})$.*

Proof. For part (i), see Koshevoy and Mosler (1995b), for part (ii), Koshevoy and Mosler (1995a), the part (iii) follows from the part (i). □

Both relative dilation and directional relative dilation are multivariate extensions of the usual univariate Lorenz ordering, i.e. the ordering of Lorenz curves. $\preceq_{dirr}$ has been named the *multivariate Lorenz order* in Mosler (1994); see also Koshevoy and Mosler (1995a). If we compare empirical distributions with the same number, say n, of support points in $\mathbb{R}^d$, dilation and directional dilation correspond to majorization and directional majorization of $n \times d$ matrices; see Marshall and Olkin (1979, ch. 15).

4 The multivariate distance-Gini index

The definition of the univariate Gini mean difference (4) has the following multivariate generalization.

Definition 4.1 *For $F \in \mathcal{F}^d$ the* distance-Gini mean difference *is*

$$M_D(F) = \frac{1}{2d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} ||x - y||\, dF(x)\, dF(y) \tag{9}$$

where $||\cdot||$ denotes the Euclidean distance in $\mathbb{R}^d$. $R_D(F) = M_D(\widetilde{F})$ is the distance-Gini index.

In the case of an empirical distribution function F_A, we get

$$M_D(F_A) = \frac{1}{2dn^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \left(\sum_{s=1}^{d} (a_{is} - a_{js})^2 \right)^{\frac{1}{2}}, \tag{10}$$

$$R_D(F_A) = \frac{1}{2dn^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \left(\sum_{s=1}^{d} \frac{(a_{is} - a_{js})^2}{\overline{a}_s^2} \right)^{\frac{1}{2}}. \tag{11}$$

Several properties of the distance–Gini mean difference and the distance–Gini index follow easily from the definitions. Recall that, for $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$, we denote $F_{\cdot\beta}(x_1, \ldots, x_d) = F(x_1\beta_1, \ldots, x_d\beta_d)$ and $F_{+\beta}(x_1, \ldots, x_d) = F(x_1 + \beta_1, \ldots, x_d + \beta_d)$.

Proposition 4.1 *For all* $F \in \mathcal{F}^d$,
(i) $0 \leq M_D(F)$,
(ii) $M_D(F) = 0$ *if and only if* F *is a one-point distribution.*
(iii) $M_D(F_{+\beta}) = M_D(F)$ *for all* $\beta_1, \ldots, \beta_d$.
(iv) M_D *is continuous w.r.t weak convergence of distributions.*

Proposition 4.2 *For all* $F \in \mathcal{F}_0^d$,
(i) $0 \leq R_D(F)$.
(ii) $R_D(F) = 0$ *if and only if* F *is a one-point distribution.*
(iii) $R_D(F_{\cdot\beta}) = R_D(F)$ *for all* $\beta_1, \ldots, \beta_d > 0$.
(iv) R_D *is continuous w.r.t weak convergence of distributions.*

Proposition 4.2(iii) says that R_D is *vector scale invariant,* while Proposition 4.1(iii) states that M_D is *translation invariant.* Regarding upper bounds we have the following result.

Theorem 4.1 *For* $F \in \mathcal{F}_+^d$, *the following inequalities hold.*

$$M_D(F) < \frac{1}{d} \sum_{j=1}^{d} \mu_j(F), \qquad R_D(F) < 1,$$

and the bounds are sharp.

We will prove the theorem at the end of this Section. Before we consider a property which is desirable for every index of multivariate disparity. It says that, if to a distribution in d attributes a $(d+1)$-th attribute is added which does not vary in the population, then the disparity index remains essentially unchanged: It multiplies by a factor which depends only on d.

Definition 4.2 (Ceteris paribus property) *Let* J^d *be a real valued function which is defined on a subset* $\mathcal{D}^d$ *of* $\mathcal{F}^d$, $d \in \mathbb{N}$. *We say that* $J^d, d \in \mathbb{N}$, *has the* ceteris paribus *property if*

$$J^{d+1}(F \otimes E_{\xi_0}) = \gamma(d) J^d(F) \quad \text{for all } F \in \mathcal{D}^d, \xi_0 \in \mathbb{R}, d \in \mathbb{N}. \tag{12}$$

Here E_{ξ_0} *denotes the univariate one-point distribution at* ξ_0, *and* $\gamma(d)$ *is a constant for every* d.

Theorem 4.2 M_D *and* R_D *have the* ceteris paribus *property with*

$$\gamma(d) = \frac{d}{d+1}.$$

The proof is obvious from the definition of M_D.

Theorem 4.3 *Let dp denote the rotation invariant area element on the sphere S^{d-1}, $d \geq 2$. There holds*

$$M_D(F) = \frac{, (\frac{d+1}{2})}{4d\,\pi^{\frac{d-1}{2}}} \int_{p\in S^{d-1}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |u-v|\; dF(u,p)\, dF(v,p) dp, \tag{13}$$

$$R_D(F) = \frac{, (\frac{d+1}{2})}{4d\,\pi^{\frac{d-1}{2}}} \int_{p\in S^{d-1}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |u-v|\; d\widetilde{F}(u,p)\, d\widetilde{F}(v,p) dp. \tag{14}$$

Proof. We use the following formula by Helgason (1980, Lemma 7.2). For every $z \in \mathbb{R}^d$ and $k > 0$ holds

$$\int_{p\in S^{d-1}} |zp^T|^k dp = \frac{2\pi^{\frac{d-1}{2}}, (\frac{k+1}{2})}{, (\frac{d+k}{2})}\; ||z||^k. \tag{15}$$

From this formula with $k = 1$, we conclude that

$$\begin{aligned}
M_D(F) &= \frac{1}{2d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} ||x-y|| dF(x)\, dF(y) \\
&= \frac{1}{2d} \frac{, (\frac{d+1}{2})}{2\pi^{\frac{d-1}{2}}} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} (\int_{p\in S^{d-1}} |xp^T - yp^T|\; dp)\; dF(x)\, dF(y) \\
&= \frac{1}{2d} \frac{, (\frac{d+1}{2})}{2\pi^{\frac{d-1}{2}}} \int_{p\in S^{d-1}} (\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |xp^T - yp^T|\; dF(x)\, dF(y)) dp \\
&= \frac{, (\frac{d+1}{2})}{4d\,\pi^{\frac{d-1}{2}}} \int_{p\in S^{d-1}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |u-v|\; dF(u,p)\, dF(v,p) dp.
\end{aligned} \tag{16}$$

This proves (13). The result for R_D follows immediately with $\widetilde{F}$ in place of F. □

Recall that the area of S^{d-1} equals $2\pi^{\frac{d}{2}}/, (\frac{d}{2})$. Equation (13) in Theorem 4.3 says that the distance-Gini mean difference M_D is a constant times the average, over all directions p in the sphere, of the Gini indices of all univariate distribution functions $F(\cdot, p)$,

$$M_D(F) = \frac{, (\frac{d+1}{2})\pi^{\frac{1}{2}}}{d\;, (\frac{d}{2})} \left[\frac{, (\frac{d}{2})}{2\pi^{\frac{d}{2}}} \int_{p\in S^{d-1}} M(F(\cdot,p)) dp \right], \tag{17}$$

and similarly for $R_D(F)$. Recall, that the Euler Gamma-function $, (s) = \int_0^\infty t^{s-1} e^{-t} dt$ has the following properties: $\sqrt{\pi} = , (\frac{1}{2})$ and $, (s+1) = s, (s)$; and

the Euler Beta-function $\mathrm{B}(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$ is equal to , (a), (b)/, $(a+b)$. Therefore,

$$\frac{,\,(\frac{d+1}{2})\pi^{\frac{1}{2}}}{d\,,\,(\frac{d}{2})} = \frac{,\,(\frac{d+1}{2}),\,(\frac{1}{2})}{2\,,\,(\frac{d+2}{2})} = \frac{\mathrm{B}(\frac{d+1}{2},\frac{1}{2})}{2}.$$

By the mean value theorem we conclude:

Corollary 4.1 *For every F there exist some p and $\tilde{p} \in S^{d-1}$ such that*

$$\begin{aligned} M_D(F) &= \frac{B(\frac{d+1}{2},\frac{1}{2})}{2}\, M(F(\cdot,p)) \quad \textit{and} \\ R_D(F) &= \frac{B(\frac{d+1}{2},\frac{1}{2})}{2}\, R(\widetilde{F}(\cdot,\tilde{p})). \end{aligned}$$

The corollary says that, for every distribution F, there are directions p and $\tilde{p}$ which reflect the dependence structure of F, i.e. the interplay between the attributes, for the Gini mean difference and the Gini index, respectively.

Remark 4.1 $M_D(F)$ is related to the lift zonoid $\widehat{Z}(F)$ as follows. For $p = (p_1,\ldots,p_d) \in S^{d-1}$, let pr_p denote the projection of $\mathbb{R}^{d+1}$ on the two dimensioned plane which is spanned by the vectors $(1,0,\ldots,0)$ and $(0,p_1,\ldots,p_d)$. Then, for $z = (z_0,z_1,\ldots,z_d) \in \mathbb{R}^{d+1}$, we get $pr_p(z) = (z_0, \sum z_i p_i)$ with respect to this base. The projection of the lift zonoid by pr_p equals the lift zonoid of $F(\cdot,p)$ (Koshevoy and Mosler 1995b). So, we can state that $M_D(F)$ is $\mathrm{B}(\frac{d+1}{2},\frac{1}{2})/2$ times the average area of these two dimensioned projections of the lift zonoid. The following proof of Theorem 4.1 uses this fact.

Proof of Theorem 4.1. For $F \in \mathcal{F}_+^d$, holds $\widehat{Z}(F) \subset [0,1] \times [\mathbf{0}, \mu(F)]$. Therefore, in view of Remark 4.1, holds

$$\begin{aligned} M(F(\cdot,p)) &= \frac{1}{2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |u-v|\; dF(u,p)\, dF(v,p) \\ &= \mathrm{V}_2(pr_p(\widehat{Z}(F)) \le \mathrm{V}_2(pr_p([0,1]\times[\mathbf{0},\mu(F)])). \end{aligned} \tag{18}$$

Recall that V_2 denotes the two-dimensional volume. Thus, by (17),

$$M_D(F) \le \frac{,\,(\frac{d+1}{2})}{2d\,\pi^{\frac{d-1}{2}}} \int_{p\in S^{d-1}} \mathrm{V}_2(pr_p([0,1]\times[\mathbf{0},\mu(F)]))dp.$$

Given $p \in S^{d-1}$, the projection $pr_p([0,1]\times[\mathbf{0},\mu(F)])$ is a rectangle whose edges have length 1 and $\sum|\mu_j(F)p_j|$ and whose area amounts to $\sum|\mu_j(F)p_j|$. Therefore,

$$M_D(F) \le \frac{,\,(\frac{d+1}{2})}{2d\,\pi^{\frac{d-1}{2}}} \int_{p\in S^{d-1}} \sum_{j=1}^{d} |\mu_j(F)p_j| dp$$

$$= \frac{,\left(\frac{d+1}{2}\right)}{2d\,\pi^{\frac{d-1}{2}}} \sum_{j=1}^{d} \int_{p\in S^{d-1}} |\mu_j(F)p_j|dp. \qquad (19)$$

In view of (15), we get

$$\frac{,\left(\frac{d+1}{2}\right)}{2\,\pi^{\frac{d-1}{2}}} \int_{p\in S^{d-1}} |\mu_j(F)p_j|dp = ||(0,\ldots,0,\mu_j(F),0,\ldots,0)|| = \mu_j(F). \qquad (20)$$

Thus, (19) and (20) yield $M_D(F) \leq \frac{1}{d}\sum_j \mu_j(F)$.

The strict inequality is due to the fact that every lift zonoid is contained in the $(d+1)$–dimensional rectangle $[0,1]\times[\mathbf{0},\mu(F)]$, but the latter is no lift zonoid.

It is easily seen that the upper bound $d^{-1}\sum_j \mu_j(F)$ cannot be improved. For example, consider the $n\times d$ matrix A whose j-th row is $(0,\ldots,0,\, n\mu_j(F),0,\ldots,0)$, $j = 1,\ldots,d$, while other rows are $(0,\ldots,0)$. Then $\lim_{n\to\infty} M_D(F_A) = \lim_{n\to\infty} \frac{n-d}{n}\, d^{-1}\,\sum_j \mu_j(F)$, which shows that $d^{-1}\sum_j \mu_j(F)$ is the least upper bound for the mean distance–Gini mean difference.

The least upper bound for the distance–Gini index is established by passing from F to $\widetilde{F}$. Recall that $\mu_j(\widetilde{F}) = 1$ for $j = 1,\ldots,d$. □

5 The multivariate volume–Gini index

Here we start with the definition of the univariate Gini index as twice the area between the Lorenz curve and the diagonal and extend it to the multivariate case. Given $F\in\mathcal{F}^d$, let $X, X_1,\ldots,X_d$ be independent random vectors each of which is distributed according to F. Q denotes the $(d+1)\times(d+1)$ matrix having rows $(1,X),(1,X_1),\ldots,(1,X_d)$, and $E|\det Q|$ is the expectation of the modulus of its determinant. The term $(d\,!)^{-1}E|\det Q|$ was called a multivariate Gini index by Wilks (1960); see Oja (1983) and Giovagnoli and Wynn (1995). Oja (1983) has interpreted it via the average volume of random simplexes with vertices $X, X_1,\ldots,X_d$. The following theorem shows that $((d+1)!)^{-1}E|\det Q|$ equals the volume of the lift-zonoid of F.

Theorem 5.1 *Let F be a given distribution function in $\mathbb{R}^d$. Let $X, X_1,\ldots,X_d$ be independent random vectors each of which is distributed according to F, and let Q denote the $(d+1)\times(d+1)$ matrix having rows $(1,X),(1,X_1),\ldots,(1,X_d)$. Then*

$$\mathrm{V}_{d+1}(\widehat{Z}(F)) = \frac{1}{(d+1)!}E|\det Q|.$$

Proof. Zonoids are limits of zonotopes. Recall, that a zonotope in $\mathbb{R}^k$ is the Minkowski sum of line segments, say

$$\overline{\mathbf{0}, y_1} + \ldots + \overline{\mathbf{0}, y_n} \subset \mathbb{R}^k \quad \text{with some given } y_i \in \mathbb{R}^k. \tag{21}$$

It has volume (see, e.g., Shephard 1974)

$$\sum_{1 \le i_1 < \ldots i_k \le n} |\det[y_{i_1}, \ldots, y_{i_k}]|. \tag{22}$$

For a given F, there exists a sequence F_ν, $\nu \in \mathbb{N}$, of distribution functions with finite supports in $\mathbb{R}^d_+$ which converges weakly to F, i.e., $\lim_\nu \int g dF_\nu = \int g dF$ for every continuous and bounded function $g : \mathbb{R}^d \to \mathbb{R}$. Due to the continuity of zonoids with respect to weak convergence (Bolker 1969), we have $\lim_\nu \delta(\widehat{Z}(F_\nu), \widehat{Z}(F)) = 0$, where δ is the Hausdorff distance. The volume is a continuous function with respect to the Hausdorff distance. Therefore, $\mathrm{V}_{d+1}(\widehat{Z}(F)) = \lim_\nu \mathrm{V}_{d+1}(\widehat{Z}(F_\nu))$. Each volume $\mathrm{V}_{d+1}(\widehat{Z}(F_\nu))$ can be calculated by the formula (22). Let F_ν have atoms at $x_1, \ldots, x_m$ with probabilities $q_1, \ldots, q_m$. Then $\widehat{Z}(F_\nu) = \overline{\mathbf{0}, (q_1, q_1 x_1)} + \ldots + \overline{\mathbf{0}, (q_m, q_m x_m)}$. Hence

$$\begin{aligned}
\mathrm{V}_{d+1}(\widehat{Z}(F_\nu)) &= \sum_{1 \le i_1 < \ldots i_{d+1} \le m} |\det[(q_{i_1}, q_{i_1} x_{i_1}), \ldots, (q_{i_{d+1}}, q_{i_{d+1}} x_{i_{d+1}})]| \\
&= \frac{1}{(d+1)!} \sum_{i_1, \ldots, i_{d+1} = 1}^{m} q_{i_1} \cdot \ldots \cdot q_{i_{d+1}} |\det[(1, x_{i_1}), \ldots, (1, x_{i_{d+1}})]| \\
&= \frac{1}{(d+1)!}\, E|\det Q_{F_\nu}|.
\end{aligned}$$

This completes the proof. □

However, the volume of a lift-zonoid equals zero rather often, also if F is no one-point distribution. Observe, that if the vectors $x_1, \ldots, x_n$ are linearly dependent, then the volume of the zonotope in (21) equals zero. Thus, whenever the support of F is contained in a linear subspace of $\mathbb{R}^{d+1}$ with dimension less than $d+1$, then the volume of the lift zonoid is zero. In the case of an empirical distribution F, if, e.g., one of the attributes is equally distributed in the population, or if two attributes have the same distribution then $\mathrm{V}_{d+1}(\widehat{Z}(F)) = 0$.

The volume of the Lorenz zonoid is given by the following formula.

$$\mathrm{V}_{d+1}(LZ(F)) = \frac{1}{\prod_{j=1}^{d} |\mu_j|} \mathrm{V}_{d+1}(\widehat{Z}(F)). \tag{23}$$

In Mosler (1994) the $(d+1)$-dimensional volume of $LZ(F)$ has been introduced as a multivariate Gini index, called the *Gini zonoid index*. Although this index shows

a number of useful properties (boundedness between 0 and 1, 0 at one-point distributions, vector scale invariance, weak monotonicity with multivariate dilations), it may be zero also at distributions which are not concentrated at one point. To avoid this drawback of the Gini zonoid index, we propose the following definition. Let $C^d = \{(z_0, z_1, \ldots, z_d) \in \mathbb{R}^{d+1} : z_0 = 0,\ 0 \le z_s \le 1,\ s = 1, \ldots, d\}$, which is a d-dimensional cube in $\mathbb{R}^{d+1}$. Instead of the volume of the lift zonoid, we use the volume of the lift zonoid 'expanded' by this cube.

Definition 5.1 *The* volume-Gini mean difference *is defined by*

$$M_V(F) = \frac{1}{2^d - 1}\left(\mathrm{V}_{d+1}(\widehat{Z}(F) + C^d) - 1\right). \tag{24}$$

$R_V(F) = M_V(\widetilde{F})$ *is the* volume-Gini index.

Let $d = 1$. Since $|x - y| = |\det(\begin{smallmatrix} 1 & 1 \\ x & y \end{smallmatrix})|$ we conclude that the the distance–Gini index and the volume-Gini index are the same. This observation allows us to extend Proposition 2.1(ii) to an arbitrary distribution $F \in \mathcal{F}_0^1$, dropping the assumption that $F(0) = 0$.

The choice of the constant $1/(2^d - 1)$ in (24) will be explained in the following theorem. We need some notations: For a nonempty subset $K \subset \{1, \ldots, d\}$, $F^{(K)}$ denotes the marginal distribution with respect to the coordinates indexed by K.

Theorem 5.2

$$M_V(F) = \frac{1}{2^d - 1} \sum_{\emptyset \ne K \subset \{1,\ldots,d\}} \mathrm{V}_{|K|+1}(\widehat{Z}(F^{(K)})), \tag{25}$$

$$R_V(F) = \frac{1}{2^d - 1} \sum_{\emptyset \ne K \subset \{1,\ldots,d\}} \mathrm{V}_{|K|+1}(\widehat{Z}(\widetilde{F}^{(K)})). \tag{26}$$

Note that Formula (2), M_V for empirical distributions, follows from (22) and (25).

Remark 5.1 By Equation (25), the volume-Gini mean difference is the average of the volumes of projections of the lift zonoid on coordinate subspaces. They are spanned by $(1, 0, \ldots, 0)$ and $(0, \mathbf{e}_r)$, $r \in K$, $K \subset \{1, \ldots, d\}$. Here $\mathbf{e}_r$ is the r-th coordinate unit vector in $\mathbb{R}^d$.

Proof of Theorem 5.2. We will prove (25) for an empirical distribution F. Then an approximation argument yields (25) for a general distribution. (26) obviously

follows from (25). Let F have atoms at $x_1, \ldots, x_m$ in $\mathbb{R}^d$ with probabilities $q_1, \ldots, q_m$. Then

$$\widehat{Z}(F) + C^d = \overline{\mathbf{0}, (q_1, q_1 x_1)} + \ldots + \overline{\mathbf{0}, (q_m, q_m x_m)} + \sum_{s=1}^{d} \overline{\mathbf{0}, (0, \mathbf{e}_s)}.$$

Hence, by (22)

$$\begin{aligned} \mathrm{V}_{d+1}(\widehat{Z}(F) + C^d) \quad = \quad & \sum_{1 \le i_1 < \ldots i_{d+1} \le m} |\det[(q_{i_1}, q_{i_1} x_{i_1}), \ldots, (q_{i_{d+1}}, q_{i_{d+1}} x_{i_{d+1}})]| \\ & + \sum_{l=1}^{d-1} \sum_{1 \le i_1 < \ldots i_{d+1-l} \le m} \sum_{1 \le s_1 < \ldots s_l \le d} \\ & |\det[(q_{i_1}, q_{i_1} x_{i_1}), \ldots, (q_{i_{d+1-l}}, q_{i_{d+1-l}} x_{i_{d+1-m}}), (0, \mathbf{e}_{s_1}), \ldots, (0, \mathbf{e}_{s_l})]| \\ & + \sum_{i=1}^{m} |\det[(q_i, q_i x_i), (0, \mathbf{e}_1), \ldots, (0, \mathbf{e}_d)]|. \end{aligned}$$

Let $1 \le l \le d-1$ and $1 \le s_1 < \ldots s_l \le d$ be fixed, $K = \{1, \ldots, d\} \setminus \{s_1, \ldots, s_l\}$. Then we have

$$\begin{aligned} & \mathrm{V}_{|K|+1}(\widehat{Z}(F^{(K)})) = \qquad (27) \\ & \sum_{1 \le i_1 < \ldots i_{d+1-l} \le m} \det[(q_{i_1}, q_{i_1} x_{i_1}), \ldots, (q_{i_{d+1-l}}, q_{i_{d+1-l}} x_{i_{d+1-l}}), (0, \mathbf{e}_{s_1}), \ldots, (0, \mathbf{e}_{s_l})]|. \end{aligned}$$

In view of $q_1 + \ldots + q_m = 1$,

$$\sum_{i=1}^{m} |\det[(q_i, q_i x_i)(0, \mathbf{e}_1), \ldots, (0, \mathbf{e}_d)]| = 1. \qquad (28)$$

(27) and (28) yield (25). □

The following three theorems establish properties of R_V and M_V.

Proposition 5.1 *For all $F \in \mathcal{F}^d$,*
(i) $0 \le R_V(F)$,
(ii) $R_V(F) = 0$ *if and only if F is a one-point distribution,*
(iii) $R_V(F_{.\beta}) = R_V(F)$ *for all $\beta_1, \ldots, \beta_d > 0$.*
(iv) R_V *is continuous w.r.t. weak convergence of distributions.*
(v) *If $F \in \mathcal{F}^d_+$, then $R_V(F) < 1$ and the bound is sharp.*

Proof. (i) The volume is a nonnegative function.
(ii): If F is a one-point distribution, then, for every K, $\widehat{Z}(\widetilde{F}^{(K)})$ is the main diagonal

of the unit hypercube in $\mathbb{R}^{\{|K|+1\}}$ and has volume zero. Therefore $R_V(F) = 0$. If F is no one-point distribution, at least one of its univariate marginals, say $F^{(j^*)}$, is the same. Then the univariate Gini index $R(F^{(j^*)})$ is positive. Since $\mathrm{V}_2(\widehat{Z}(\widetilde{F}^{(j^*)})) = R(F^{(j^*)})$, at least one summand in (26) does not vanish, and therefore $R_V(F) > 0$.
(iii): The vector scale invariance is obvious from the definition of $R_V(F)$, since it is based on the relative distribution $\widetilde{F}$ only.
(iv) follows from Theorem 7.1 in Koshevoy and Mosler (1995b).
(v): For every K, $\widehat{Z}(\widetilde{F}^{(K)})$ is contained in the unit hypercube of $\mathbb{R}^{|K|+1}$, hence $0 \le \mathrm{V}_{|K|+1}(\widehat{Z}(\widetilde{F}^{(K)})) < 1$, and, by (26), $0 \le R_V(F) < 1$. It is easily seen that the upper bound 1 cannot be improved. For example, consider the distribution $F(x) = \prod_{i=1}^{d} F_i(x_i)$ where $F_i(x_i) = 0$ if $x_i < 0$, $F_i(x_i) = (n-1)/n$ if $0 \le x_i < 1$, $F_i(x_i) = 1$ if $x_i \ge 1$. Then $R_V(F) \to 1$, for $n \to \infty$. □

Proposition 5.2 *For all $F \in \mathcal{F}^d$,*
(i) $0 \le M_V(F)$,
(ii) $M_V(F) = 0$ *if and only if F is a one-point distribution,*
(iii) $M_V(F_{+\beta}) = M_V(F)$ *for all $\beta_1, \ldots, \beta_d$.*
(iv) M_V *is continuous w.r.t. weak convergence of distributions.*
(v) *If $F \in \mathcal{F}^d_+$, then*

$$M_V(F) < \frac{1}{2^d - 1} \sum_{\emptyset \neq K \subset \{1,\ldots,d\}} \prod_{i \in K} \mu_i \le \frac{1}{2^d - 1} \Big((\max_i \mu_i + 1)^d - 1 \Big)$$

and the first inequality cannot be improved.

The proof is similar to that of Proposition 5.1.

Theorem 5.3 M_V *and* R_V *have the* ceteris paribus *property with*

$$\gamma(d) = \frac{2^d - 1}{2^{d+1} - 1}.$$

Proof. It is easily seen, that $\mathrm{V}_{|K|+1}(\widehat{Z}((F \otimes E_\xi)^{(K)}) = 0$ if $d+1 \in K$. If $d+1 \notin K$ then $F^{(K)} = (F \otimes E_\xi)^{(K)}$. This and (25) yield the proposition. □

6 Consistency with multivariate dilations

The univariate Gini index respects dilation and Lorenz order. We will show that our distance-Gini and volume-Gini indices do the same for properly defined extensions of these orderings.

Proposition 6.1 *The following implications holds*
(i) $F \preceq G \quad \Rightarrow \quad F \preceq_r G \quad \Rightarrow \quad F \preceq_{dirr} G$.
(ii) $F \preceq G \quad \Rightarrow \quad F \preceq_a G \quad \Rightarrow \quad F \preceq_{dira} G$.
(iii) $F \preceq G \quad \Rightarrow \quad F \preceq_{dir} G \quad \Rightarrow \quad F \preceq_{dirr} G$ *and* $F \preceq_{dira} G$.
(iv) $F \preceq_{dirr} G \quad \Rightarrow \quad R(F(\cdot,p)) \leq R(G(\cdot,p))$ *for all* $p \in S^{d-1}$.

Proof. A standard characterization of dilation says that $F \preceq G$ if and only if $\int \phi(x)dF(x) \leq \int \phi(x)dG(x)$ holds for all convex functions $\mathbb{R}^d \to \mathbb{R}$; see, e.g., the references in Mosler (1994). Further, $F \preceq G$ implies $\mu(F) = \mu(G)$.
(i): Assume $F \preceq G$, and let $\phi : \mathbb{R}^d \to \mathbb{R}$ be convex. Then, with $(\mu_1, \ldots, \mu_d) = \mu(F) = \mu(G)$, the function $x \mapsto \phi(\frac{x_1}{\mu_1}, \ldots, \frac{x_d}{\mu_d})$ is convex, too. We conclude

$$\begin{aligned} \int \phi(x)d\widetilde{F}(x) &= \int \phi\left(\frac{x_1}{\mu_1}, \ldots, \frac{x_d}{\mu_d}\right)dF(x) \\ &\leq \int \phi\left(\frac{x_1}{\mu_1}, \ldots, \frac{x_d}{\mu_d}\right)dG(x) = \int \phi(x)d\widetilde{G}(x). \end{aligned}$$

Therefore $F \preceq_r G$. Now assume that $F \preceq_r G$. Let $p \in S^{d-1}$, $\psi : \mathbb{R} \to \mathbb{R}$ convex. Then the function $x \mapsto \psi(xp^T)$ is convex, and from $F \preceq_r G$ follows that

$$\begin{aligned} \int \psi(u)d\widetilde{F}(u,p) &= \int \psi(xp^T)d\widetilde{F}(x) \\ &\leq \int \psi(xp^T)d\widetilde{G}(x) = \int \psi(u)d\widetilde{G}(u,p), \end{aligned}$$

hence $F \preceq_{dirr} G$.
(ii): The proof is similar to that of (i).
(iii): Dilation implies directional dilation. The rest follows from parts (i) and (ii) with $d = 1$.
(iv): If $F \preceq_{dirr} G$ and $p \in S^{d-1}$, then $F(\cdot,p)$ is smaller than $G(\cdot,p)$ in relative dilation (= usual Lorenz order). As the usual Gini index is consistent with Lorenz order, we conclude (iv). □

Note that, besides the implications given in Proposition 6.1, in general no other implications hold between the various multivariate dilations.

Proposition 6.2 (i) $\preceq, \preceq_{dir}$ *are partial orders (reflexive, transitive, antisymmetric) in* $\mathcal{F}^d$.
(ii) $\preceq_r$ *and* $\preceq_{dirr}$ *are preorders (reflexive, transitive) in* $\mathcal{F}_0^d$.
(iii) $\preceq_a$ *and* $\preceq_{dira}$ *are preorders (reflexive, transitive) in* $\mathcal{F}^d$.

Note that the preorders $\preceq_r$, $\preceq_{dirr}$, $\preceq_a$ and $\preceq_{dira}$ are also antisymmetric when applied to the proper factor space.

Proof. (i): The antisymmetry of $\preceq_{dir}$ is proven in Koshevoy and Mosler (1995 b). The antisymmetry of $\preceq$ follows from the antisymmetry of $\preceq_{dir}$ and Proposition 6.1. (ii) and (iii) follow from (i) and Proposition 6.1. □

Theorem 6.1 *The distance-Gini index R_D and the volume-Gini index R_V are strictly increasing with*
(i) *dilation,*
(ii) *directional dilation,*
(iii) *relative dilation,*
(iv) *directional relative dilation.*

Proof. In view of Proposition 6.1, only (iv) has to be shown. Suppose $F \preceq_{dirr} G$, hence $R(F(\cdot,p)) \leq R(G(\cdot,p))$ for all $p \in S^{d-1}$. Then

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |u-v|\, d\widetilde{F}(u,p)\, d\widetilde{F}(v,p) \leq \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |u-v|\, d\widetilde{G}(u,p)\, d\widetilde{G}(v,p)$$

for all p. Therefore,

$$\int_{p\in S^{d-1}}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |u-v|\, d\widetilde{F}(u,p)\, d\widetilde{F}(v,p)dp$$

$$\leq \int_{p\in S^{d-1}}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |u-v|\, d\widetilde{G}(u,p)\, d\widetilde{G}(v,p)dp$$

for all p. This yields, according to Proposition 4.3, $R_D(F) \leq R_D(G)$. The result for R_V follows immediately from Theorems 3.1, 5.2 and the following Proposition 6.3. That the indices are strictly increasing is seen from Theorems 3.1, 5.2 and the following Theorem 6.2. □

Proposition 6.3 (Koshevoy and Mosler (1995a)) *Let $F \preceq_{dirr} G$. Then $F^{(K)} \preceq_{dirr} G^{(K)}$ for all K, $\vee \neq K \subset \{1,\ldots,d\}$.*

Theorem 6.2 (Koshevoy and Mosler (1995b)) *$\widehat{Z}(F) = \widehat{Z}(G)$ iff $F = G$.*

For the distance-Gini and the volume-Gini mean differences, we have an analogous theorem.

Theorem 6.3 *The distance-Gini mean difference M_D and volume-Gini mean difference M_V are strictly increasing with*
(i) *dilation,*
(ii) *directional dilation,*
(iii) *absolute dilation,*
(iv) *directional absolute dilation.*

Proof. Proofs of (i) and (ii) are similar to those of (i) and (ii) in Theorem 6.1. (iii) and (iv) follow from Propositions 4.1 and 5.2 respectively.

7 Conclusions

We have presented two different approaches to extend the usual Gini index and Gini mean difference to the multivariate case. Both approaches preserve important properties of the univariate notions, are increasing with proper multivariate dilations and have the *ceteris paribus* property. The distance–Gini index and the volume–Gini index of a given empirical distribution are easily calculated, but the latter needs more computation time. A computer program, written in GAUSS, can be obtained from the authors.

Many other multivariate definitions are possible. A popular approach is to use the arithmetic mean, M_S resp. R_S, of the univariate indices,

$$M_S(F_A) = \frac{1}{2n^2 d} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{s=1}^{d} |a_{is} - a_{js}|, \tag{29}$$

$$R_S(F_A) = \frac{1}{2n^2 d} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{s=1}^{d} |\frac{a_{is}}{\overline{a}_i} - \frac{a_{js}}{\overline{a}_j}|. \tag{30}$$

This is tantamount to employing the L_1 distance instead of the Euclidean distance in our distance–Gini notions. It can be shown that always $R_D(F) \leq R_S(F)$ and $R_V(F) \leq R_S(F)$ hold. But this approach, as the index depends on the marginals only, does not reflect the dependency structure of the underlying distribution.

To illustrate and contrast our notions, we calculate them for R. A. Fisher's Iris data (Fisher 1936). The data include the measurements of four attributes, sepal length and width and petal length and width, of fifty plants for each of three types of Iris, *Iris setosa, Iris versicolor* and *Iris virginica*. The data have been used to test the

hypothesis that Iris versicolor is a polyploid hybrid of the two other species which is related to the fact that Iris setosa is a diploid species with 38 chromosomes, Iris virginica is a tetraploid, and Iris versicolor is a hexaploid with 108 chromosomes.

	Iris setosa	Iris versicolor	Iris virginica
R_D	0.08536007	0.12217668	0.14415565
R_V	0.042062259	0.067639891	0.083820681
R_S	0.13663000	0.20862000	0.24658000
$R[1]$	0.19620000	0.29000000	0.35136000
$R[2]$	0.20624000	0.17508000	0.17508000
$R[3]$	0.092760000	0.25992000	0.30616000
$R[4]$	0.051320000	0.10948000	0.15372000

Table 1. The multivariate Gini indices R_D, R_V and R_S for three types of Iris; data from Fisher (1939). For further contrast, the univariate Gini index $R[k]$ is given for each attribute k, $k = 1, 2, 3, 4$.

As we can see from the Table, the four attributes are most variable at different types of Iris, as measured by their univariate Gini indices. E.g., the first attribute, petal length, varies most with Iris virginica, while the second attribute, petal width, has its maximum Gini index with Iris setosa. But our three multivariate Gini indices, R_D, R_V, and R_S, order the variability of the three samples in the same way,

$$\text{Iris setosa} < \text{Iris versicolor} < \text{Iris virginica}.$$

Note, however, that no two of these multivariate indices are order equivalent in general.

Under the assumptions that (1) a hybrid has an intermediate number of chromosomes compared to its origins and (2) that a higher number of chromosomes implies more variability, we may conclude that all three multivariate Gini indices back the hypothesis that Iris versicolor is a hybrid of the two others species.

Acknowledgements

We thank Stephan Erkel for his comments on a previous version and Ulrich Casser for writing the computer program and calculating the numerical example.

References

BOLKER, E.D. (1969). A class of convex bodies. *Transactions of the American Mathematical Society* **145**, 323–346.

FISHER, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7/2**, 179–188.

GIORGI, G.M. (1990). Bibliographic portrait of the Gini concentration ratio. *Metron* **48**, 183–221.

GIORGI, G.M. (1992). *Il rapporto di concentrazione di Gini.* Siena: Libreria Editrice Ticci.

GIOVAGNOLI, A., & WYNN, H.P. (1995). Multivariate dispersion orderings. *Statistics and Probability Letters* **22**, 325–332.

HELGASON, S. (1980). The Radon transform. *Progress in Mathematics* **5**. Boston, Stuttgart: Birkhäuser.

KOSHEVOY, G.A., & MOSLER, K. (1995a). The Lorenz zonoid of a multivariate distribution. Mimeo.

KOSHEVOY, G.A., & MOSLER, K. (1995b). A geometrical approach to compare the variability of random vectors. *Discussion Papers in Statistics and Quantitative Economics* **66**, UniBw Hamburg.

MARSHALL, A.W., & OLKIN, I. (1979). *Inequalities: Theory of Majorization and Its Applications*. New York: Academic Press.

MOSLER, K. (1994). Majorization in economic disparity measures. *Linear Algebra and Its Applications* **199**, 91–114.

NYGÅRD, F., & SANDSTRÖM, A. (1981). *Measuring Income Inequality*. Stockholm: Almqvist and Wiksell.

OJA, H. (1983). Descriptive statistics for multivariate distributions. *Statistics and Probability Letters* **1**, 327–332.

SHEPHARD, G.C. (1974). Combinatorial properties of associated zonotopes. *Canadian J. of Mathematics* **26**, 302–321.

TAGUCHI, T. (1981). On a multiple Gini's coefficient and some concentrative regressions. *Metron* (1981), 69–98.

TORGERSEN, E. (1991). *Comparison of Statistical Experiments.* Cambridge University Press, Cambridge, Massachusets.

WILKS, S.S. (1960). Multidimensional statistical scatter. In: I. Olkin et al. eds. *Contributions to Probability and Statistics in Honor of Harold Hotelling.* Stanford, California, 486–503.