The robustness of the generalized Gini index

The robustness of the generalized Gini index

S. Settepanella[1], A. Terni[2], M. Franciosi,[3], and L. Li[4]

[1]*Department of Economics and Statistics, Torino University, Italy, simona.settepanella@unito.it*
[2]*Department of Mathematics, Pisa University, Pisa, Italy, terniale@gmail.com*
[3]*Department of Mathematics, Pisa University, Pisa, Italy, marco.franciosi@unipi.it*
[4]*School of Economics, Guangzhou College of Commerce, Guangzhou, China, lile7k@gmail.com*

Abstract

In this paper we introduce a map Φ, which we call *zonoid map*, from the space of all non-negative, finite Borel measures on $\mathbb{R}^n$ with finite first moment to the space of zonoids of $\mathbb{R}^n$. This map, connecting Borel measure theory with zonoids theory, allows to slightly generalize the Gini volume introduced, in the context of Industrial Economics, by Dosi et al. (2016). This volume, based on the geometric notion of zonoid, is introduced as a measure of heterogeneity among firms in an industry and it turned out to be a quite interesting index as it is a multi-dimensional generalization of the well known and broadly used Gini index.
By exploiting the mathematical context offered by our definition, we prove the continuity of the map Φ which, in turn, allows to prove the validity of a SLLN type theorem for our generalized Gini index and, hence, for the Gini volume. Both results, the continuity of Φ and the SLLN theorem, are particularly useful when dealing with a huge amount of multi-dimensional data.

MSC 2010 Classfication: 28B05; 28A78

keywords: Gini Index, zonoid, empirical distribution, Hausdorff metric

1 Introduction

Many problems in the social and system sciences are naturally multivariate and cannot be easily represented with a continuous or parametric approach.

An example is the economical production theory that studies and represents the determinant factors driving production process dynamics. An industry is defined as a set of firms operating within the same sector and we can think about firm productivity as the "ability" to turn inputs into outputs.
The classic approach in production theory is based on a number of assumptions regarding firm behaviour and firm production possibilities, in particular the profit maximization and the cost minimization assumption. Following these assumptions, an *ad hoc* parametrized family of production functions is introduced to estimate a number of economical indices and to assess both, the productivity and the efficiency of a firm. Production functions satisfy, in addition, certain topological properties such as convexity and continuity, thus implying that firms with similar technologies will adopt analogous production techniques or, equivalently, that firms tend to be *homogeneous*.
Despite these assumptions, a growing availability of longitudinal microdata at firm-level has evidenced the fundamental role of heterogeneity in all relevant aspects regarding firms production activity, thus suggesting a switch from a continuous/parametric approach (which seems to be inadequate in presence of wide asymmetries) to a discrete/nonparametric point of view. Here geometry and geometric measure theory come into help.
To evidence the fragilities of the classic theory, Hildenbrand (1981) adopted a different perspective, by considering the empirical distribution induced by a set $X = \{y_n\}_{n=1,\dots,N} \subset \mathbb{R}^{m+1}_+$ of firms composing the industry (see Section 4 for details), and introducing a geometric approach, the *zonoid representation*. Geometrically, a zonoid is a centrally symmetric, compact, convex set of the euclidean space which is induced by a Borel measure with finite expectation. In particular, the zonoid induced by the empirical distribution of a given industry is a convex polytope which is called a *zonotope*. Zonotopes can also be written as a sum of line segments, in addition, they are dense in the space of zonoids with respect to the topology induced by the Hausdorff metric.
More recently, Dosi et al. (2016) (see also Dosi et al. (2021)) adopted Hildenbrand's construction to assess the rate of productivity and technological change of a given industry both on the microeconomic point of view (i.e. firm-level productivity) and on the macroeconomic point of view (i.e. aggregate productivity). Moreover a measure of heterogeneity of the industry, called the *Gini volume*, is introduced. The above approach relies entirely on the geometry of the zonotope induced by the empirical distribution of the industry and it is highly nonparametric. The Gini volume can also be seen as a measure of concentration of the empirical distribution. Indeed it is nothing else than a multi-dimensional generalization of the well known Gini index broadly used in social sciences and economics as measure of statistical concentration (see Remark 4.11).
The aim of this paper is to look at the Gini volume in a slightly more general

mathematical context than the one in Dosi et al. (2016). This broader setting includes tools of measure theory and geometric properties of zonoids. It allows to generalize the definition of Gini volume to a broader class of measures and to prove the validity of a strong law of large numbers (SLLN for short) result for this generalized index. This turns out to be very useful when dealing with huge number of high dimensional data.
We introduce the zonoid map $\Phi\colon \mathcal{M}^n \to \mathcal{Z}^n$ from the space $\mathcal{M}^n$ of all non-negative, finite Borel measures on $\mathbb{R}^n$ with finite first moment to the space $\mathcal{Z}^n$ of zonoids of $\mathbb{R}^n$. This is possible thanks to the dual aspect, provided by the zonoid representation, between the theory of Borel measures with finite first moment and the geometry of convex bodies. Such map turns out to be continuous and allows to prove the validity of a SLLN type theorem for the Gini volume. More precisely, we prove the continuity of Φ on the subspace of Borel probability measures with support on a compact $K \subset \mathbb{R}^n$ (see Proposition 2.2). In turn, Proposition 2.2 provides the key ingredient to prove the main result of this paper, Theorem 4.13.
Another interesting consequence of the continuity of Φ is that every "discrete" distribution μ can be substituted by a suitable "continuous" distribution ν in such a way that the zonoid $Z(\nu) = \Phi(\nu)$ is a good approximation of $Z(\mu) = \Phi(\mu)$ at any desirable degree. This seems to suggest that a very large but finite data-set can be approximated with a continuous distribution, which may simplify the analysis without a great loss of information. This will be object of further studies.
Moreover, from the continuity of the map Φ, we can deduce a notion of robustness for the Gini volume. Indeed small changes in the value of the distribution induced by a concrete data-set X (e.g. of technological data), lead to a small change in the related zonoid and, consequently, in the Gini volume. In turn, this robustness allows to improve the computational aspect of the method by considering random samples instead of the whole data-set in order to compute the Gini volume.
In conclusion, it is worth to remark that our approach is in the same spirit of the one used in Koshevoy and Mosler (1997). Their generalizations of the Gini index and Gini mean difference to the multi-dimensional case adopt very similar mathematical techniques. For example, Corollary 3.3 and an analogous of Theorem 4.13 apply to their generalization too. On the other hand, their indices assess different quantities from applied point of view. For instance, the way the volume of the Zonotope is normalized in order to give rise to the two indices is different. In particular, while we generalize the normalization introduced by Dosi et al. (2016), a normalization chosen for its applied meaning (as explained in Subsection 4.2), their approach uses the concept of *lift Zonoid* and it is useful to straightforwardly apply several mathematical results.
For further applications of zonoid theory to other branches of economics, such as finance and stochastic processes, we refer to Molchanov and Schmutz (2011), Molchanov et al. (2014).

Zonoids can also be defined as the expectation of a random segment. For a thorough investigation of this different approach we refer to Mosler (2002) and, for a complete introduction to the general theory of random sets and its applications to econometrics, to Molchanov (2018) and Molchanov and Molinari (2018).
The paper is organized as follows. In Section 2 we introduce basic notions and preliminary results. In Section 3 we define the empirical distribution and the empirical zonoid and prove that a SLLN theorem holds. In Section 4 we investigate the zonotope approach in production theory proposed in Hildenbrand (1981), we generalize the Gini volume introduced in Dosi et al. (2016) and we present a SLLN result for this new generalized Gini index. In Section 5 we present applications of our result and in Section 6 our conclusions.

2 Notations and preliminary results

A *zonoid* is a convex body of $\mathbb{R}^n$ (i.e. it is compact and convex) which is centrally symmetric and contains the origin. A *zonotope* is a Minkowski sum of a finite number of line segments. In particular a zonoid is a polytope if and only if it is a zonotope. In this section we recall their relation with measure theory. We mainly refer to Bolker (1969), Billingsley (1968), and Mosler (2002). For a more detailed presentation of the content of this and the following Section in the context of this paper see Terni (2019).

2.1 An introduction to zonoids

Let $\mathcal{M}^n$ be the set of all non-negative, finite Borel measures μ on $\mathbb{R}^n$ (with respect to the euclidean topology) whose first moment

$$m(\mu) = \int_{\mathbb{R}^n} x \; d\mu(x)$$

is well defined (here the integration is made component-wise). For every $\mu \in \mathcal{M}^n$, the *zonoid* associated to the measure μ is the set

$$Z(\mu) = \left\{ \int_{\mathbb{R}^n} \phi(x) \cdot x \; d\mu(x) \middle| \; \phi \colon \mathbb{R}^n \to [0,1] \text{ measurable} \right\} \subseteq \mathbb{R}^n.$$

It can be considered as a geometric representation of the underlying measure: indeed, if we denote with $\mathcal{B}^n$ the class of Borel subsets of $\mathbb{R}^n$, then the zonoid $Z(\mu)$ can be seen as the closure of the convex hull of the image of the map

$$F \colon \mathcal{B}^n \to \mathbb{R}^n \; ; \; F(B) = \int_B x \; d\mu(x).$$

The zonoid $Z(\mu)$ is centrally symmetric about $\frac{1}{2}m(\mu)$ (sometimes we may also refer to $m(\mu)$ as the *mean* or the *gravity center* of the distribution).
On the functional point of view, if we denote by $\mathcal{Z}^n$ the set of zonoids of $\mathbb{R}^n$ we can consider the map

$$\Phi\colon \mathcal{M}^n \to \mathcal{Z}^n \ ; \ \Phi(\mu) = Z(\mu),$$

which we call the *zonoid map*. The zonoid map satisfies the following properties:

1. it is a homomorphism of semigroups: $Z(\mu + \nu) = Z(\mu) + Z(\nu)$ for every μ, $\nu \in \mathcal{M}^n$, where the sum on the right-hand side of the equality is the Minkowski sum;

2. it is positively homogeneous: for every $\alpha > 0$ we have $Z(\alpha\mu) = \alpha Z(\mu)$;

3. it is linearly equivariant: for every linear map $L\colon \mathbb{R}^n \to \mathbb{R}^k$ we have $L(Z(\mu)) = Z(L_*\mu)$, where $L_*\mu$ is the push-forward measure of μ with respect to L. In particular, the linear image of a zonoid is a zonoid.

In addition, the zonoid map is clearly surjective but on the other hand it is not injective, since every zonoid is induced by a measure with support contained in the unitary sphere S^{n-1} (for a proof, see Bolker (1969)).

2.2 Zonotopes and zonoids

First of all note that a zonoid is a zonotope if and only if it is induced by a finite atomic measure, i.e. a measure with finite support (cfr. Bolker (1969)).
Now, let $\mathcal{K}^n$ be the set of convex bodies of $\mathbb{R}^n$. It is a classical result that if we equip $\mathcal{K}^n$ with the Hausdorff distance

$$d_H(K, L) = \min\left\{\epsilon \geq 0 |\ K \subseteq L + \epsilon \cdot B^n,\ L \subseteq K + \epsilon \cdot B^n\right\},$$

where B^n is the unit ball in $\mathbb{R}^n$, then $(\mathcal{K}^n, d_H)$ is a complete, sequentially compact metric space.
Since the set of polytopes is dense in $\mathcal{K}^n$ with respect to the topology induced by the Hausdorff distance, the subset of zonotopes is dense in $\mathcal{Z}^n \subseteq \mathcal{K}^n$. That is, every zonoid can be arbitrarily approximated (in the Hausdorff metric) by a zonotope, which has both a geometrical and combinatorial nature (see Bolker (1969) for the proof and the geometrical characterization of a zonotope and Ziegler (1995) for the combinatorial aspects). It is worth remarking that in combinatorial geometry there is an identification between zonotopes and arrangements of hyperplanes, although we won't deal with these aspects of the theory. Figure 1 displays a zonotope generated by 4 line segments in $\mathbb{R}^3$.

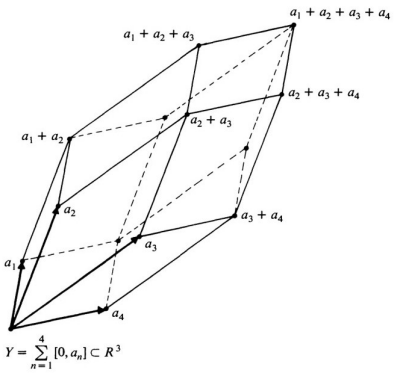

Figure 1: Zonotope generated by 4 line segments.

2.3 Continuity of the zonoid map

In this subsection and in the rest of the paper we will deal with the space $\mathcal{P}^n(K)$ of Borel probability measures with support contained in K and equipped with the topology induced by the weak convergence. Here K is either a compact subset of $\mathbb{R}^n$, i.e. $K \in \mathcal{C}^n$, or the non-negative octant $\mathbb{R}^n_+$ or the whole space $\mathbb{R}^n$. In the latter case we simply write $\mathcal{P}^n$ instead of $\mathcal{P}^n(\mathbb{R}^n)$.
We recall that a sequence $(\mu_n)_{n\in\mathbb{N}} \subset \mathcal{P}^n(K)$ is said to converge *weakly* to $\mu \in \mathcal{P}^n(K)$ if

$$\lim_{n\to\infty} \int_K f \, d\mu_n = \int_K f \, d\mu$$

for every real-valued, continuous and bounded function f defined on K. In this case we write $\mu_n \Rightarrow \mu$.
Let $\mathcal{P}^n_1(K) := \mathcal{P}^n(K) \cap \mathcal{M}^n$ be the space of probability measures with finite first moment and whose support is contained in K. A family of measures $(\mu_i)_{i\in I}$ in $\mathcal{P}^n_1(K)$ is *uniformly integrable* if

$$\lim_{\beta\to\infty} \sup_{i\in I} \int_{\|x\|\geq\beta} \| \, x \, \| \;\; d\mu_i(x) = 0.$$

The following Theorem, corollary of a more general result related to lift zonoids[1] (see Section 2.4 of Mosler (2002)), holds.

Theorem 2.1. *Let $(\mu_k)_{k\in\mathbb{N}}$, $\mu \in \mathcal{P}^n_1(K)$. If (μ_k) is uniformly integrable and $\mu_k \Rightarrow \mu$, then $Z(\mu_k) \xrightarrow{d_H} Z(\mu)$.*

[1]For a more detailed discussion on lift zonoids in the context of this work we refer the interested reader to Terni (2019).

Note that we have the equality $\mathcal{P}_1^n(K) = \mathcal{P}^n(K)$ when K is compact. In particular, a family of measures $(\mu_i)_{i\in I}$ in $\mathcal{P}^n(K)$ is always uniformly integrable when K is compact. Hence, as a corollary of Theorem 2.1, we have the following Proposition.

Proposition 2.2 (Continuity on compact sets)**.** *For every $K \in \mathcal{C}^n$, the zonoid map*

$$\Phi\colon \mathcal{P}^n(K) \to \mathcal{Z}^n \ ; \ \Phi(\mu) = Z(\mu)$$

is continuous.

Proof. Every family of measures with support contained in a compact set is uniformly integrable. Hence, by Theorem 2.1 the map Φ is a sequentially continuous map between two metric spaces, in particular it is a continuous map. □

As aforementioned, beside the case in which K is a compact set, it is of common interest the case in which K coincides with $\mathbb{R}^n_+$.
Set $\mathcal{P}_1^+ = \mathcal{P}_1^n(\mathbb{R}^n_+)$. We are interested in describing another sufficient condition, beside uniform integrability, that a family $(\mu_k)_{k\in\mathbb{N}}$ of measures in $\mathcal{P}_1^+$ needs to satisfy in order to obtain a convergence result. With this aim we recall that a sequence $(\mu_k)_{k\in\mathbb{N}} \subset \mathcal{P}_1 = \mathcal{P}_1^n(\mathbb{R}^n)$ is said to be convergent *in mean* to $\mu \in \mathcal{P}_1$ (write $\mu_k \xrightarrow{\mathcal{M}} \mu$) if it converges weakly to μ and the sequence $(m(\mu_k))$ converges to $m(\mu)$ for $k \to \infty$. Hildenbrand (1981) proved the following result.

Theorem 2.3. *Given $(\mu_k)_{k\in\mathbb{N}} \subset \mathcal{P}_1^+$ and $\mu \in \mathcal{P}_1^+$, then $\mu_k \xrightarrow{\mathcal{M}} \mu$ implies $Z(\mu_k) \xrightarrow{d_H} Z(\mu)$.*

Remark that for any K compact subset of $\mathbb{R}^n$, a sequence $(\mu_k)_{k\in\mathbb{N}} \subset \mathcal{P}(K)$ is convergent in mean to $\mu \in \mathcal{P}(K)$ if and only if it is weakly convergent to μ.
Before to move to the next section, we briefly recall here that a fundamental example of Borel probability distribution on $\mathbb{R}^n$ is the *Dirac measure* $\delta_x \in \mathcal{P}^n, x \in \mathbb{R}^n$, defined as:

$$\delta_x(B) = \begin{cases} 0, & \text{if } x \notin B \\ 1, & \text{if } x \in B \end{cases}$$

for every B Borelian subset of $\mathbb{R}^n$.
Clearly, the support of the Dirac measure δ_x coincides with the singleton $\{x\}$. In addition, the space of *atomic probability measures* (i.e. those distributions with finite support) coincides with the space

$$\mathcal{Q}^n = \left\{ \sum_{i=1}^N \alpha_i \delta_{x_i} \in \mathcal{P}^n\colon \ N \in \mathbb{N},\ x_1, \dots, x_N \in \mathbb{R}^n,\ \sum_{i=1}^N \alpha_i = 1,\ \alpha_i \in [0,1] \right\} \quad .$$

This is the space of convex combinations of Dirac measures which is a dense subset of $\mathcal{P}^n$ with respect to the topology induced by the weak convergence (further details can be found in Billingsley (1968)).
The Dirac measure plays an important role in the next and in the last section of this paper.

3 Zonoids related to empirical distributions

We begin with the following Definition.

Definition 3.1. Let $X = \{y_k\}_{k=1,\dots,N} \subset \mathbb{R}^n$ be a finite set. The *empirical distribution* of X is the Borel measure

$$\widehat{\mu} = \frac{1}{N} \sum_{k=1}^{N} \delta_{y_k},$$

the zonoid related to the empirical distribution $Z(\widehat{\mu})$ is the *empirical zonoid.*

As noticed in Subsection 2.2, since $\widehat{\mu}$ is a measure with finite support then the induced empirical zonoid $Z(\widehat{\mu})$ is indeed a zonotope.
In many application contexts, the empirical distribution is induced by a data-set X of technological data which are subject to errors of various kind. Hence, it is desirable that a small change in the distribution should lead only to a small change in the related zonoid or, equivalently, that the map Φ should satisfy a continuity result. This is quite useful when one needs to rely on samples, for instance when the collection of technological data (e.g. the production activity of an industry in several countries) is time consuming and costly. In this respect, in Proposition 2.2 we have already stated a continuity result for zonoids in the compact case. An analogous result can be stated for the non compact case $\mathcal{P}_1^+$. The following version of the Glivenko-Cantelli Theorem for separable metric spaces holds (see Varadarajan (1958)).

Theorem 3.2. *Let (E, d) be a separable metric space and $X_1, X_2, \dots$ be independent E-valued random variables with distribution μ (we consider on E the σ-field of Borelian subsets). Let $\widehat{\mu}_N$ be the empirical measure*

$$\widehat{\mu}_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{X_i} \quad ,$$

then we have $\widehat{\mu}_N \Rightarrow \mu$ for $N \to \infty$ with probability 1.

Notice that Theorem 3.2 implies that the empirical zonoid which is derived from a large sample of the true distribution μ will yield a good approximation of $Z(\mu)$. A consequence of Theorem 3.2 and Theorem 2.3 is the following Corollary.

Corollary 3.3. *Let $X_1, X_2, \ldots$ be independent $\mathbb{R}^n_+$-valued random variables with distribution $\mu \in \mathcal{P}_1^+$. Let $\widehat{\mu}_N$ be the empirical measure*

$$\widehat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i} \quad ,$$

then we have

$$Z(\widehat{\mu}_N) \xrightarrow{d_H} Z(\mu)$$

with probability 1.

Proof. The usual law of large numbers implies $m(\widehat{\mu}_N) \xrightarrow{\|\cdot\|} m(\mu)$ with probability 1, hence we can combine it with Theorem 3.2 to conclude that $\widehat{\mu}_N \xrightarrow{\mathcal{M}} \mu$ with probability 1 and thus the thesis follows by Theorem 2.3. □

To conclude we remark that Corollary 3.3 can actually be extended to $X_1, X_2, \ldots$ independent $\mathbb{R}^n$-valued random variables with distribution $\mu \in \mathcal{P}_1$ (see Mosler (2002)).

4 A generalization of the Gini index.

In recent years, a wide literature based upon empirical analyses has robustly evidenced the permeating presence of heterogeneity in all relevant aspects of the dynamics of production processes. Recently Dosi et al. (2016) introduced the *Gini Volume*, a new non parametric index to assess the degree of heterogeneity of an industry. Their construction is based on the paper Hildenbrand (1981), in which the author applies the theory of zonoids to the one of industrial production. In this section we recall the definition of such index, we provide a slight generalization by means of the zonoid representation and we prove the validity of a SLLN type result.

4.1 The zonotope approach

Hildenbrand (1981) suggested a geometrical representation of a given industry. Such representation is highly nonparametric and it is based upon observed production activity, that is, every industry is represented as a set

$$X = \{y_n\}_{n=1,\ldots,N} \subset \mathbb{R}^{m+1}_+,$$

where:

- N is the number of productive units (i.e. the firms) making up the industry;
- every point y_n is called the *observed* production activity of the n-th firm;
- the first m coordinates of y_n represent the input quantities adopted by the n-th firm and the last coordinate is the output quantity produced under the period of observation (we say we are in the m-input, 1-output case)[2].

Let $X = \{y_n\}_{n=1,\dots,N} \subset \mathbb{R}_+^{m+1}$ be a fixed set which represents a given industry. Hildenbrand (1981) defines the *production set* of the n-th firm as the line segment

$$[0, y_n]\,.$$

The *size* of the n-th firm is the euclidean norm of the vector $\overrightarrow{0y_n}$, $\parallel y_n \parallel$.
Notice that the definition of production set corresponds, roughly speaking, to the assumption that each firm doesn't change its production activity under the period of observation, thus it can be seen as a first order approximation of the problem. In Hildenbrand (1981) there is a geometric representation of the industry X from the aggregate point of view.

Definition 4.1. The *short-run total production set* of the industry X is the Minkowski sum of the production set of each firm, that is, the zonotope

$$Z = \sum_{n=1}^{N} [0, y_n].$$

Consider the empirical measure of the industry X, that is, the measure

$$\widehat{\mu} = \frac{1}{N} \sum_{n=1}^{N} \delta_{y_n}.$$

We recall that $\widehat{\mu}$ is a probability measure with finite support, hence it is an atomic probability with finite mean and we have $\widehat{\mu} \in \mathcal{P}_1^+$. As noted by Hildenbrand, for every Borelian set B the quantity $100 \cdot \widehat{\mu}(B)$ can be seen as the percentage of production units having their characteristics in the set B.

Definition 4.2. The *short-run mean production set* of the industry X is the zonoid $Z(\widehat{\mu})$, where $\widehat{\mu}$ is the empirical distribution of X.

The term "mean" adopted in the above definition follows from the observation that $Z(\widehat{\mu})$ is an homothetic copy of the short-run total production set Z. Indeed we have

$$Z = N \cdot Z(\widehat{\mu}).$$

[2]We replaced $\mathbb{R}^n$ with $\mathbb{R}^{m+1}$ to be consistent with the notation in Hildenbrand (1981) and Dosi et al. (2016).

Remark 4.3. As a convex body, every zonoid $Z(\mu)$ is uniquely determined by its *support function*, defined as follows:

$$\psi_\mu \colon \mathbb{R}^n \to \mathbb{R} \; ; \; \psi_\mu(\xi) = \sup \{ \langle x, \xi \rangle |\; x \in Z(\widehat{\mu}) \} \, .$$

It is an interesting fact that in Hildenbrand (1981), an economic interpretation of the support function of $Z(\widehat{\mu})$ is given: if we write $\xi = (-\xi_1, \ldots, -\xi_m, \xi_{m+1}) \in \mathbb{R}^{m+1}$, then the quantity $\psi_{\widehat{\mu}}(\xi) = \sup \{ \langle x, \xi \rangle |\; x \in Z(\widehat{\mu}) \}$ can be considered as the maximum mean profit with respect to the price system ξ subject to the technological restrictions defined by the mean production set $Z(\widehat{\mu})$.

Building by Hildenbrand's work, Dosi et al. (2016) introduced a new framework to study the rate and direction of technical change and to assess the firm level heterogeneity, which we are now going to examine.

4.2 Heterogeneity and Gini volume

Empirical evidence reports a wide and persistent heterogeneity across firms operating in the same industry, thus the phenomenon requires attention.
Intuitively, heterogeneity can be associated in mathematical statistics to the variance, namely it measures how much the industry is far from being homogeneous or, equivalently, how much the various productive units differ from the "mean" productive unit.

Definition 4.4. Let $X = \{y_n\}_{n=1,\ldots,N} \subset \mathbb{R}^{m+1}_+$ be an industry and let Z be the related short-run total production set. The *total production activity* is the sum

$$\Sigma_Z = \sum_{n=1}^{N} y_n \in Z.$$

Geometrically, the line segment $d_Z := [0, \Sigma_Z]$ is the main diagonal of the zonotope Z and it seems to be a good candidate to represent the "'mean" productive technology of the industry: indeed we have

$$\frac{\Sigma_Z}{N} = m\left(\widehat{\mu}\right),$$

where $m\left(\widehat{\mu}\right)$ is the expectation of the empirical measure $\widehat{\mu}$ related to the industry (i.e. the set) X.
For a better visualization, let us analyse two limit cases, one the opposite of the other.

- **Maximal homogeneity:** every production set lies on the line spanned by the main diagonal d_Z. This corresponds to the situation where every production activity adopts the same productive technology and any two of them only differ by their intensities (i.e. their size). In this case, we have $Z = d_Z$, which is a zonotope with null volume;
- **Maximal heterogeneity:** production sets are represented by segments on positive semi-axis and the zonotope Z is a parallelotope in $\mathbb{R}^{m+1}$ with diagonal d_Z. This case has to be regarded as a limit case: indeed, production sets on positive semi-axis would imply that there are firms with either nonzero inputs and zero output or nonzero output and zero inputs, which is quite absurd.

Building from these two cases, Dosi et al. (2016) defined the following index as a candidate measure of heterogeneity.

Definition 4.5. The *Gini volume* for the short run total production set Z induced by the industry X is the ratio

$$G(Z) = \frac{V_{m+1}(Z)}{V_{m+1}(P_Z)} \in \mathbb{R},$$

where P_Z is the $(m+1)$-dimensional parallelotope

$$P_Z := \left\{ z \in \mathbb{R}^{m+1} \colon \ 0 \leq z \leq \sum_{i=1}^{N} y_n = \Sigma_Z \right\}.$$

Observe that the Gini volume does not depend on the units of measure or the number of firms, thus it allows comparisons across space and time. In addition, we have the inequality

$$0 \leq G(Z) \leq 1,$$

where the minimum is attained at the maximal homogeneity case and the maximum is attained in the maximal heterogeneity case.

Remark 4.6. Clearly, the inequality $N \geq m+1$ must be satisfied, otherwise the Gini volume would be null (observe that in applications the number N is usually large). When $N \geq m+1$, then we have the equality

$$V_{m+1}(Z) = \sum_{i \in I} |\Delta_i|,$$

where $I = \{i = (i_1, \ldots, i_{m+1}) \in \mathbb{R}^{m+1} \mid 1 \leq i_1 < \cdots < i_{m+1} \leq N\}$ and Δ_i is the determinant of the matrix whose rows are the vectors $\left\{y_{i_1}, \ldots, y_{i_{m+1}}\right\}$. On the other hand, we have

$$V_{m+1}(P_Z) = \Pi_{i=1}^{m+1} \langle \Sigma_Z, e_i \rangle,$$

where $\{e_i\}_{i=1,\ldots,m+1}$ is the canonical basis and $\langle\ ,\ \rangle$ is the standard scalar product.

The following continuity result on the Gini volume holds.

Theorem 4.7. *Let $\mathcal{Z}_+^{m+1}$ be the space of zonotopes Z that are contained in $\mathbb{R}_+^{m+1}$ and verify $V_{m+1}(P_Z) \neq 0$. Then the Gini volume, seen as a real-valued function defined on $\mathcal{Z}_+^{m+1}$ equipped with the topology induced by the Hausdorff metric, is continuous.*

In order to prove this theorem we need the following Lemma (see Schneider (2013)).

Lemma 4.8. *The volume functional V_{m+1} is continuous on the space of convex bodies in $\mathbb{R}^{m+1}$ with respect to the Hausdorff metric.*

Proof of Theorem 4.7. Since the volume functional is continuous by Lemma 4.8, the only thing left to prove is the continuity of the map

$$Z \mapsto P_Z.$$

Indeed, the function is also uniformly continuous, in fact for every couple of zonotopes Z, Z' with $d_H(Z, Z') \leq \epsilon$ we have

$$Z \subseteq Z' + \epsilon \cdot B^{m+1} \subseteq P_{Z'} + \epsilon \cdot B^{m+1},$$

hence the inclusion

$$P_Z \subseteq P_{Z'} + \epsilon \cdot B^{m+1}$$

follows easily from the definition of P_Z. Clearly we can exchange the roles of Z and Z' to get the inequality

$$d_H(P_Z, P_{Z'}) \leq \epsilon.$$

□

The Gini volume defined above can be expressed in terms of the empirical distribution $\widehat{\mu}$ of the set X as showed in the following Remark.

Remark 4.9. Note that, for every $\mu \in \mathcal{P}_1^+$, the associated zonoid $Z(\mu)$ is contained in the $m+1$-dimensional parallelotope

$$P(\mu) := \left\{ z \in \mathbb{R}^{m+1} \colon \ 0 \leq z \leq m(\mu) \right\}, \tag{1}$$

where $\leq$ is applied component by component. In this respect we have the equality

$$G(Z) = \frac{V_{m+1}(Z(\widehat{\mu}))}{V_{m+1}(P(\widehat{\mu}))} = G\left(Z\left(\widehat{\mu}\right)\right),$$

which can be easily deduced from the relations $Z = N \cdot Z(\widehat{\mu})$ and $P_Z = N \cdot P(\widehat{\mu})$. In particular, we have $V_{m+1}(P_Z) \neq 0$ if and only if the expectation $m(\widehat{\mu}) \in \mathbb{R}_+^{m+1}$ is a vector with strictly positive coordinates.

4.3 A generalized Gini index and its robustness

Remark 4.9 suggests an extension of the Gini volume definition to the set of zonoids induced by $\mathcal{P}_1^+$.

Definition 4.10. Let $\mu \in \mathcal{P}_1^+$ be a Borel distribution such that $m(\mu)$ is a vector with strictly positive coordinates. The *generalized Gini index* related to μ is the ratio

$$G\left(Z\left(\mu\right)\right) = \frac{V_{m+1}(Z(\mu))}{V_{m+1}(P(\mu))},$$

where $P(\mu)$ is the parallelotope defined in Remark 4.9.

In the following Remark we show how the generalized Gini index defined above is, indeed, a generalization of the Gini index.

Remark 4.11. Let $\mu \in \mathcal{P}_1^1$ be a univariate probability distribution with support contained in $\mathbb{R}_+$ and such that $m(\mu) \neq 0$ (equivalently $m(\mu) > 0$). Consider the *lifted measure* induced by μ, that is, the bivariate probability distribution

$$\overline{\mu} = \delta_1 \otimes \mu,$$

where $\delta_1 \in \mathcal{P}_1^1$ is the Dirac measure which assigns unitary mass to the point 1. Observe that we can write $\overline{\mu} \in \mathcal{P}_1^+$ if we set $m + 1 = 2$.
In Mosler (2002) it is proved that the zonoid $Z(\overline{\mu})$ (which is also called the *lift zonoid* induced by μ) is a bidimensional convex body bordered by two curves, the *generalized Lorenz curve* and the *dual generalized Lorenz curve* induced by μ. We recall that the generalized Lorenz curve induced by the distribution μ is defined as

$$L_\mu(t) = \left(t, \int_0^t Q_\mu(s)\ ds\right)\ ,\ 0 \leq t \leq 1,$$

where $Q_\mu(s)$ is the quantile function of μ:

$$Q_\mu(s) = \inf\left\{x \in \mathbb{R}\colon\ \mu\left(]-\infty, x]\right) \geq s\right\}.$$

The dual generalized Lorenz curve is obtained by symmetrization of the generalized Lorenz curve with respect to the center of symmetry of $Z(\overline{\mu})$, that is, the point $C = \left(\frac{1}{2}, \frac{1}{2}m(\mu)\right) \in \mathbb{R}^2$. Figure 2 shows the zonoid $Z(\overline{\mu})$ and the parallelotope $P(\overline{\mu})$ when μ is the exponential distribution with parameter 1, that is, when $\mu = Exp(1)$. The generalized Lorenz curve is represented by the lower curve below the dotted line displayed in the figure, which corresponds to the segment whose endpoints are the origin and the point $(1, m(\mu))$. The dual generalized Lorenz curve is represented by the upper curve above the dotted line. On the other hand, the rectangle (the square) containing the zonoid in Figure 2 coincides

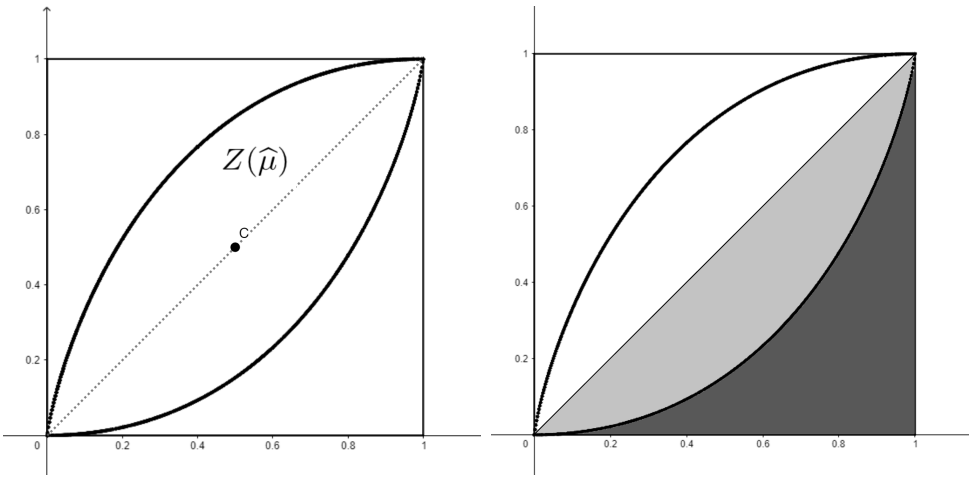

Figure 2: Lorenz curve.

with the 2-dimensional parallelotope $P(\overline{\mu})$.
On the right figure, the light grey surface represents the portion of plane between the dotted line and the generalized Lorenz curve, while the dark grey surface represents the portion of $P(\overline{\mu})$ which is situated below the generalized Lorenz curve. By a symmetry argument, we can observe that the proposed generalization in Definition 4.10 graphically coincides with the ratio between the area of the light grey surface and the area of the dark grey surface united with the light grey surface. Hence the term *generalized* Gini index in the Definition 4.10 is justified.

Let $P(\mu)$ be the parallelotope defined in Remark 4.9, the following continuity result holds.

Theorem 4.12. *Let $(\mu_k)_{k\in\mathbb{N}} \subset \mathcal{P}_1^+$ and $\mu \in \mathcal{P}_1^+$ be Borel distributions such that $V_{m+1}(P(\mu)) \neq 0$ and $V_{m+1}(P(\mu_k)) \neq 0$ for every index k. If $\mu_k \xrightarrow{\mathcal{M}} \mu$, then the sequence $G(Z(\mu_k))$ converges to $G(Z(\mu))$.*

Proof. The proof follows immediately by Theorem 2.3 and the observation that if $\mu_k \xrightarrow{\mathcal{M}} \mu$, then $P(\mu_k) \xrightarrow{d_H} P(\mu)$. □

Notice that the above Theorem applies to the index of heterogeneity proposed in Dosi et al. (2016).
The following is our final and main result, a SLLN type theorem, which may be used in a more general context, beside the production theory one.

Theorem 4.13. *Let $\mu \in \mathcal{P}_1^+$ be a Borel distribution such that the expectation $m(\mu)$ is a vector with strictly positive coordinates and let $X_1, X_2, \dots$ be independent $\mathbb{R}_+^{m+1}$-valued random variables with distribution μ. Let $\widehat{\mu}_N$ be the empirical*

measure

$$\widehat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i},$$

then the sequence $G(Z(\widehat{\mu}_N))$ *is eventually defined and it converges to* $G(Z(\mu))$ *with probability 1.*

Proof. Observe that, since we have $\mu \in \mathcal{P}_1^+$, the expectation $m(\mu)$ is a vector with strictly positive coordinates if and only if the parallelotope $P(\mu)$ has non-empty interior or, equivalently, if and only if $V_{m+1}(P(\mu)) \neq 0$. By the usual law of large numbers we have $m(\widehat{\mu}_N) \to m(\mu)$ with probability 1, hence the sequence of parallelotopes $P(\widehat{\mu}_N)$ has eventually non-empty interior and thus the index $G(Z(\widehat{\mu}_N))$ is eventually well defined almost surely. At this point, we can conclude by Theorem 3.2 and Theorem 4.12. □

5 Applications to the Gini volume

In this section we consider two examples to explain some possible applications of our results.

5.1 On the efficiency of computations via sub-samples

Recently, based on the software *Zonohedron*[3] in Dosi et al. (2016), Cococcioni et al. (forthcoming) developed a Stata[4] command to compute the Gini volume of a data-set of vectors. The computational complexity of the algorithm behind both softwares is $\mathcal{O}(N^l)$, where N and $(l+1)$ are, respectively, the number of vectors in the set considered and the dimension of the vector space. Hence, as pointed out by Cococcioni et al. (forthcoming) the use of a sub-sample can efficiently reduce the computational time.
To better estimate the extent of our results we have applied the aforementioned algorithm to the analysis of an industry composed by 1400 firms. This data sample is obtained from the data base AMADEUS[5]. We firstly considered the number of employees and the fixed assets as inputs and the turnover values as

[3] *Zonohedron* is written by Federico Ponchio and can be downloaded at `http://vcg.isti.cnr.it/~ponchio/zonohedron.php`.

[4] Stata is a general-purpose statistical software package developed by StataCorp for data manipulation, visualization, statistics, and automated reporting. Stata is very popular for empirical studies among economists.

[5] AMADEUS, a commercial database provided by Bureau van Dijk, contains balance sheets and income statements for over 21 million European firms over the period 2004-2013. We selected the 2007 data-set of an Italian industry of 1400 firms (4-digit NACE classification).

output, i.e. we considered the 3-dimensional case. It took 0.364 minutes for the Stata command to compute the Gini volume for the industry with 1400 firms. The computation time drops to 0.002 minutes when we focused on 200 firms randomly drawn from the data sample. This benefit of efficiency becomes even larger when dealing with the analysis in higher dimension. For example, if we further introduce the material cost into our analysis as a 3rd input, i.e., the dimension of the vector space is 4, the computation times for 1400 and 200 firms are, respectively, 151.844 and 0.116 minutes. In dimension 6, according to Cococcioni et al. (forthcoming), shirinking the sample size from 250 to 200 decreases the computation time by almost 12 hours.
In conclusion, considering a lower number of elements in the data-set following our continuous results, reduces drastically the time of computations of the Gini volume defined by Dosi et al. (2016).

5.2 On the accuracy of computations via sub-samples

In this subsection we address the question on the size that a sub-sample of a given data-set should have in order to get an accurate estimation of the Gini volume. We do this by means of an empirical example. Further studies are needed in order to provide a more precise theoretical answer.
Let's denote by G the Gini volume of the entire data-set and by g_j the Gini volume of the j-th round sub-sample of a fixed size. Both G and g_j are computed by means of the Stata command developed by Cococcioni et al. (forthcoming). We consider the

$$\text{standard } g_j = \frac{g_j - G}{sd_j(g_j)}$$

where $sd_j(\cdot)$ computes the standard deviation over j.
We investigate around 100 different industries[6] by fixing sub-samples of the size of $10\%, 20\%, 30\%$, and 40% for each one, re-sampling 1000 times in each case. The results are plotted in Figure 3. The majority of industries (around the 70%) behaved as represented in the left panel of Figure 3. In this case the mode provides an almost perfect approximation of G when the size of the sub-sample is the 40%. In other cases the mode of standard g_j approximates G almost perfectly already with a 10% sub-sample, as depicted in the right panel of Figure 3. In those cases the distribution of the standard g_j becomes multimodal when the sub-sample size arrives to 40%.
Those computations show that a 40% sub-sample is enough to provide a good

[6]Italian industries (4-digit NACE classification) in 2011 with the number of firms within ranges from 387 to 699 extracted from AMADEUS.

approximation of the Gini volume. Notice that if with the choice of the 40% the distribution of the standard g_j is multimodal, then a better approximation can be obtained by shrinking the size of the sub-sample.
From this example two evidences arise:

1. the original sample can be non-trivially reduced;

2. the choice of a suitable sub-sample is a problem worthy to be investigated.

Those considerations show how our theoretical result on the robustness of the generalized Gini index can be fruitfully applied to the computation of the Gini volume making it faster and, in some cases, feasible.
On the other hand there is an unexpected and interesting consequence of this example. Our robustness result could indirectly provide a new way to study the distribution of firms in an industry. In particular, it could cast a light on how the different techniques in an industry are used, which ones are the most popular and which are the most effective (over time). Indeed if we consider the industries represented in the right panel of Figure 3, it is reasonable to infer that the distribution of the firms inside those industries is rather different than the distribution of the firms inside the industries represented in the left panel. One possible explanation for this difference is that in this minority of industries, the firms distribute in clusters of homogeneous techniques. Indeed in this case if we re-sample too many firms within one cluster (still possible for random re-sampling), the g_j approximates only the Gini volume of that cluster but not necessarily the Gini volume, G, of the industry. This is consistent with the multimodal distribution of the standard g_j when the sub-sample size arrives to 40%. Hence, by re-grouping the firms which show homogeneous techniques, we could be able to identify the prominent techniques in the industry and study them (over time). Since establishing which are the most effective techniques in an industry is a problem widely studied, we believe that this finding deserves further studies.

6 Conclusions

In this paper we deal with a multi-dimensional generalization of the well known Gini index. Our generalization moved from the definition of the Gini volume provided by Dosi et al. (2016) and its first application is to the computation of this exact Gini volume. Indeed the Gini volume defined by Dosi et al. (2016) is a very useful tool to study the heterogeneity of an industry, but its computational complexity in higher dimension makes difficult to use it in the interesting case in which a large number of inputs is involved. Theorem 4.13 provides a theoretical result which can be used to reduce the sample size and hence the computational

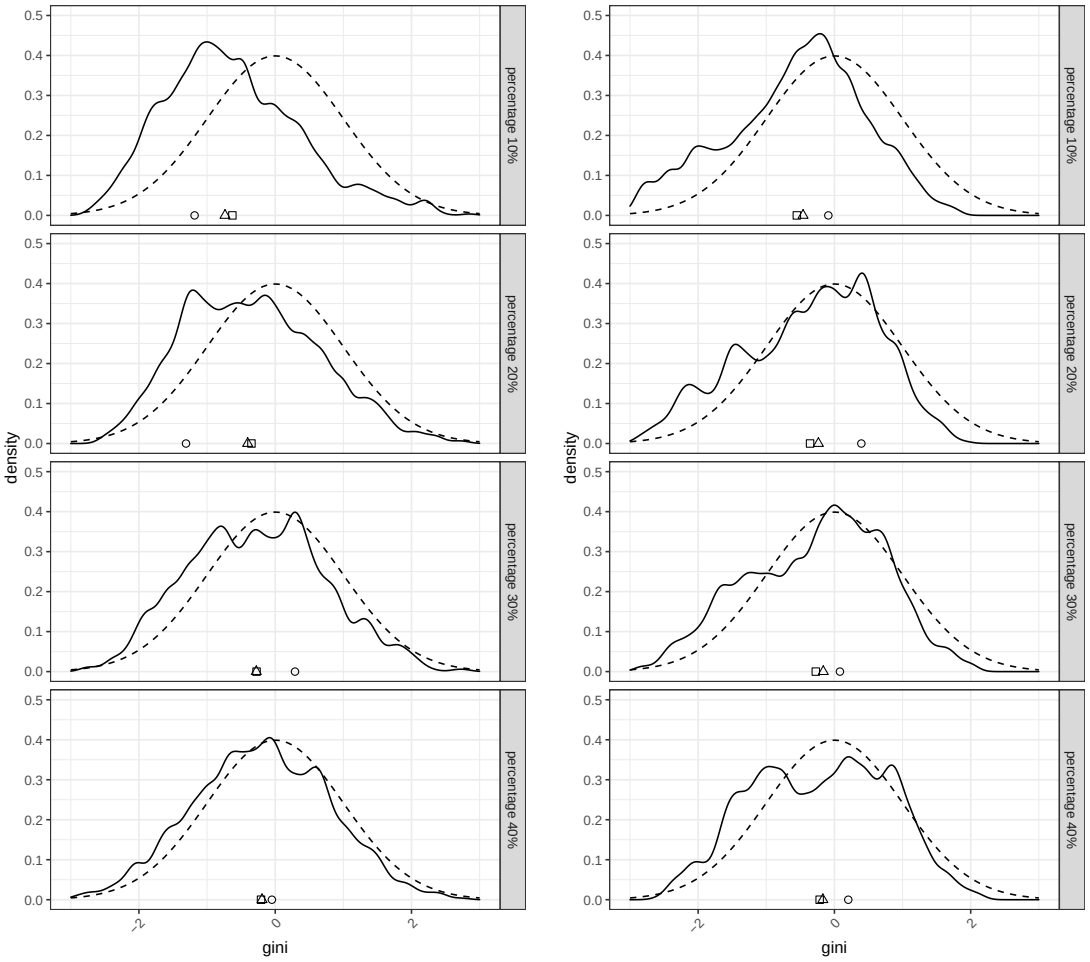

Figure 3: The kernel distributions of the standard g_j for different sub-sample sizes (solid line) compared with the standard normal distribution (dashed line). The $\triangle$, $\square$, and $\circ$ represent the mean, median, and mode of the distributions of the standard g_j.

complexity of the Gini volume. Moreover the examples studied in Subsection 5.2 show how the distribution of the firms inside an industry is a non-trivial and an interesting function to be studied. Indeed a more accurate study of those distributions could answer to interesting questions such as:

1. are there techniques which are dominant in a given industry?

2. are the dominant techniques the most efficient ones?

3. is the efficiency of the dominant techniques predictive of the future growth of the industry?

All those questions are very important in Industrial Economics and this could provide a new way to investigate them from a totally different point of view.

Compliance with Ethical Standards .
Funding: No funds, grants, or other support was received.
Disclosure of potential conflicts of interest: The authors have no conflicts of interest to declare that are relevant to the content of this article.
Informed consent: not applicable

References

P. Billingsley. *Convergence of Probability Measures*. John Wiley and Sons, New York, 1968.

E. Bolker. A class of convex bodies. *Transactions of the American Mathematical Society*, 145:323–345, 1969.

M. Cococcioni, M. Grazzi, L. Li, and F. Ponchio. A toolbox for measuring heterogeneity and eciency using zonotopes. *The Stata Journal*, forthcoming.

G. Dosi, M. Grazzi, L. Marengo, and S. Settepanella. Production theory: Accounting for firm heterogeneity and technical change. *Journal of Industrial Economics*, 4:875–907, 2016.

G. Dosi, M. Grazzi, L. Li, L. Marengo, and S. Settepanella. Productivity decomposition in heterogeneous industries. *Journal of Industrial Economics, to appear*, 69:615–652, 2021.

W. Hildenbrand. Short-run production functions based on microdata. *Econometrica*, 49:1095–1125, 1981.

G. Koshevoy and K. Mosler. Multivariate gini indices. *Journal of Multivariate Analysis*, 60:252–276, 1997.

I. Molchanov. *Theory of Random Sets*. Springer, 2018.

I. Molchanov and F. Molinari. *Random Sets in Econometrics*. Cambridge University Press, Cambridge, 2018.

I. Molchanov and M. Schmutz. Exchangeability-type properties of asset prices. *Advances in Applied Probability*, 43 (3):666–687, 2011.

I. Molchanov, M. Schmutz, and K. Stucki. Invariance properties of random vectors and stochastic processes based on the zonoid concept. *Bernoulli*, 20 (3):1210–1233, 2014.

K. Mosler. *Multivariate Dispersion, Central Regions and Depth: The Lift Zonoid Approach*. Springer-Verlag, New York, 2002.

R. Schneider. *Convex Bodies: The Brunn-Minkowski Theory, 2nd Ed.* Cambridge University Press, Cambridge, 2013.

A. Terni. *A Geometric Characterization of Borel Distributions with Applications in Nonparametric Statistics*. Master Thesis, University of Pisa, 2019.

V. Varadarajan. On the convergence of sample probability distributions. *Sankya, The Indian Journal of Statistics*, 19:23–26, 1958.

G. Ziegler. *Lectures on Polytopes*. Springer-Verlag, New York, 1995.