# A Toolbox for Measuring Heterogeneity and Efficiency using Zonotopes

Marco Cococcioni
Department of Information Engineering, University of Pisa
56122 Pisa, Italy
marco.cococcioni@unipi.it
Marco Grazzi
Department of Economic Policy, Università Cattolica del Sacro Cuore
20123 Milano, Italy
marco.grazzi@unicatt.it
Le Li
School of Economics, Guangzhou College of Commerce
511363 Guangzhou, China
lile7k@gmail.com
Federico Ponchio
Visual Computing Lab, ISTI CNR
56124 Pisa, Italy
federico.ponchio@isti.cnr.it

**Abstract.** In this work we describe a new stata command **zonotope** that, by resorting to a geometry-based approach, provides a measure of productivity that fully accounts for the existing heterogeneity across firms within the same industry. Further, the method that we propose also enables to assess the extent of multi-dimensional heterogeneity with applications to fields beyond that of production analysis only. Finally, we detail the functioning of the software to perform the related empirical analysis and we discuss the main computational issues encountered in its development.

**Keywords:** st0001, zonotope, heterogeneity measures, Stata

## 1 Introduction

In this work we present a software that allows to measure the existing heterogeneity among multi-dimensional units that can also be nested in groups at different level of aggregation. While in this paper we document the application of the methodology to the context of production analysis in economics, the possible range of applications is much wider. Within the proposed field, the method also provides a measure of productivity (both at the aggregate and at the individual level) and its change over time, that allows to relax many of the standard assumptions of the theory, which have been falsified by recent work on actual data.

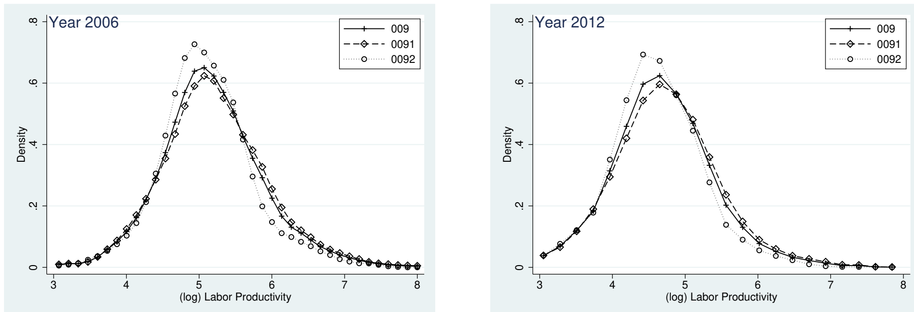
Traditionally, empirical analysis in economics has suffered from the scarcity of dis-aggregated sources of data (i.e. at the level of individual, household, enterprise, etc.), so that in the analysis of behaviors at the micro level, much was left to theoretical analysis oftentimes requiring heroically simplifying assumptions on the behavior of agents, the trade-offs they were facing, the absence of any path-dependency, etc. Nowadays, the ever growing availability of disaggregated data on business firms has revealed a much richer picture than what previously conjectured on the basis of theories alone or on aggregate industry-level data.

Firms are different along most of the dimensions typically taken into consideration by economic analyses. To provide a brief account of what is at stake, consider that firms, even within the same, narrowly defined, industry display very different levels of productivities, both in terms of labor and total factor productivities, (see among the others, Baily et al. 1992; Baldwin and Rafiquzzaman 1995; Bartelsman and Doms 2000; Disney et al. 2003; Syverson 2011; Dosi et al. 2016), the relative input intensities are much different (Dosi and Grazzi 2006) even in presence of the same relative input prices and, at least as relevant, such heterogeneities are persistent over time, i.e. if there is any selection at work, it takes much long to display its effects (Dosi 2007). The ubiquitous presence of such heterogeneity has been vividly expressed by Griliches and Mairesse (1999): "We [...] thought that one could reduce heterogeneity by going down from general mixtures as 'total manufacturing' to something more coherent, such as 'petroleum refining' or 'the manufacture of cement.' But something like Mandelbrot's fractal phenomenon seems to be at work here also: the observed variability-heterogeneity does not really decline as we cut our data finer and finer. There is a sense in which different bakeries are just as much different from each others as the steel industry is from the machinery industry."

The evidence recalled above presents several challenges to the standard theory of production and to the related empirical applications based on the notion of a "representative" firm or of an industry production function, and, of course, on the estimation of such production function itself. More in detail, the observed combinations of inputs chosen by firms appear to be quite dispersed, hardly displaying any regularity resembling a conventional isoquant. Further, although output - as expected - increases in both inputs, however this does happen in a non-monotonic way (Dosi and Grazzi 2006). Also the degree to which firms actually operate substitution among their inputs is challenged by empirical observations.

Granted that, how can one obtain measures of productivity, both at the firm and at the industry level, that do not require assumptions which are not met by data, and on the contrary, take into account such pervasive heterogeneity within industry? In this work we present a new application that enables to measure productivity in presence of such firm-level heterogeneity. In addition, the methodology that is developed also allows to provide a measure of the degree of heterogeneity of the firms making up the industry and assess its variation over time. Note that the latter feature is open to applications in a wider range of fields than production analysis only. For example, based on Zonotopes and Path Polytopes inclusion, Andreoli and Zoli (2014) frame the dissimilarity comparisons of sets of groups distributions which can be applied into multi-group comparisons of segregation, discrimination and mobility, as well as inequality evaluations.

To illustrate the phenomenon of heterogeneity and its relevance, in Figure 1




(a) Year 2006 (b) Year 2012

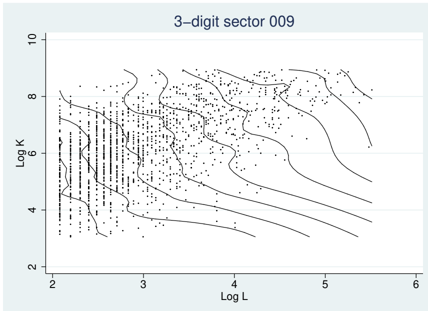
Figure 1: Empirical Distribution of (log) labor productivity in a given 3-digit sector, and in two nested 4-digit sectors.

we provide some empirical evidence focusing on the labor productivity distribution, where labor productivity is defined as the ratio of deflated turnover value over number of employees and the details on these two variables can be found in section 4.[1]

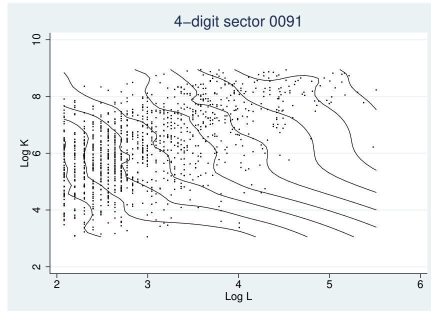
```
. use "empirical_firm_data.dta", clear
. gen LogProductivity = log(Turnover/NumberOfEmployees)
. kdensity LogProductivity if Year==2006, nograph generate(x y009)
. kdensity LogProductivity if Year==2006 & NACE=="0091", nograph generate(y0091) at(x)
. kdensity LogProductivity if Year==2006 & NACE=="0092", nograph generate(y0092) at(x)
. label var y009 "009"
. label var y0091 "0091"
. label var y0092 "0092"
. twoway connected y009 y0091 y0092 x if x < 8 & x > 3, sort lc(black black black)
lp(solid dash dot) mcolor(black black black) msymbol(plus Dh Oh) ytitle(Density) xtitle((log)
Labor Productivity) title(Year 2006, pos(10) ring(0)) xscale(range(3(1)8)) legend(col(1)
pos(1) ring(0))
```

In the left panel of Figure 1, one can appreciate that the productivity distribution of a given 3-digit sector in 2006, i.e. the solid line, is sufficiently widespread to indicate the huge productivity gap between the most productive firms and the least productive ones. Note that in the graph productivity is measured in a log scale. This makes the heterogeneity even greater. Further this heterogeneity does not disappear when focusing on more similar firms,

[1]For illustrative purpose, the following lines of code enable to reproduce Figure 1, left panel (the same holds for Figure 2a). Data and code (stata do file) to reproduce all of the evidence reported in the paper are contained in the "replication" folder. More details on the data employed in this work are reported in Section 4.




(a) 3-digit sector




(b) nested 4-digit sector

Figure 2: Contour plots of adopted techniques and output in 3- and 4-digit sectors

that is moving from the 3-digit, to the 4-digit industrial classification,[2] i.e. the dashed and dotted lines; we still observe significantly different productivity levels among firms. The persistence of heterogeneity indicates not only that heterogeneity holds when increasing the level of dis-aggregation but also that it holds over time.

As recalled above, not only firms within the same industry display very different levels of efficiency, but they also employ very diverse production techniques. Figure 2a helps to envisage this other property which emerges from the data.

```
. use "empirical_firm_data.dta", clear
. keep if (NACE == "0091" | NACE == "0092") & Year == 2006
. gen LogK = log(FixedAssets)
. gen LogL = log(NumberOfEmployees)
. gen LogOutput = log(Turnover)
. keep if LogK > 3 & LogK < 9 & LogL > 2 & LogL < 6
. bipolate LogOutput LogK LogL, collapse(mean) method(thinplatespline) smooth(0.1)
xlevels(40) ylevels(40) saving(bipolate_result, replace)
. gen LogOutput_mean = 0
. append using bipolate_result
. twoway contourline LogOutput_mean LogK LogL if LogOutput_mean > 0, levels(10) ||
(scatter LogK LogL if LogOutput_mean == 0, msize(vtiny) mcolor(black)), title("3-digit
sector 009") xtitle("Log L") ytitle("Log K")
```

Figure 2a provides a representation of production activities of firms within

[2]For statistical purpose, firms are assigned to industries according to their main activity. If a firm, as is often the case, is diversified (i.e. is active in more than one industry) it is assigned to the most important sector, the one generating more revenues. There exist several classifications of economic activities, see for instance Eurostat (2008) or United Nations (2008).

a given 3-digit industry[3] assuming the standard 2-input-1-output production, where the axes represent inputs (labor and capital are proxied by number of employees and fixed assets, respectively) and the contour line displays a constant level of output as proxied by turnover value. Thus, each firm within this industry, as one observation in our empirical data sample, can in principle be represented by one point in this contour plot. In the input plane, we plot first all such points representing the firms' labor and capital combinations from empirical data and, second, the isoquants indicating the possible combination of labor and capital corresponding to the same output level. It is easy to see that dots in the input plane are quite dispersed while the corresponding isoquants do not display any regularity resembling conventional production function (for a more in depth discussion refer to Hildenbrand 1981; Dosi and Grazzi 2006). Furthermore, the output increases with the inputs in a non-monotonic manner. To be specific, given a quantity of one input, different firms attain the same level of output with very different levels of the other input. Again, notice that, also this type of heterogeneity does not disappear when we increase the disaggregation of the industrial classification; similar phenomena can be observed in 4-digit level industries, as reported in Figure 2b. To sum up, this empirical evidence suggests that within one industry not only firms display very different levels of productivity, but also that their production techniques are much heterogeneous. As a result, it would be a great improvement for productivity measurement, and its change over time, to fully account for such apparent differences among firms. It is on this attempt that we focus in the next Section.

## 2 The Zonotope approach to production analysis

In this section we outline in brief the geometric approach to production analysis on which we rely for the proposed software packages. For a more detailed exposition we refer to Hildenbrand (1981) and Dosi et al. (2016).

The seminal work by Hildenbrand (1981) suggests an agnostic and data oriented approach which instead of estimating some aggregate production function, offers a representation of the empirical production possibility set of an industry in the short run based on actual microdata. In such a settings it is possible to represent a firm (or, for that matter, an establishment) in the input-output space. In such a way the production possibility set of any given industry is represented geometrically by the space formed by the finite sum of all the line segments linking the origin and the points representing each production unit, called a zonotope. Based on this zonotope framework, Dosi et al. (2016) move a step forward and show that by further exploiting the properties of zonotopes it is possible to obtain rigorous measures of heterogeneity and productivity without imposing on data a model like that implied by standard production functions.

Similar to Koopmans (1977); Hildenbrand (1981), and now also to Dosi et al. (2016), we denote the production activity, as representing the actual technique

[3] The message would not change picking a different one.

of production unit $i$, by a vector

$$a_i = (\alpha_{i_1}, \cdots, \alpha_{i_l}, \alpha_{i_{l+1}}) \in \mathbb{R}_+^{l+1} \tag{1}$$

which indicates that during the current period, this production unit, at its best, can produce $\alpha_{i_{l+1}}$ units of output by means of $(\alpha_{i_1}, \cdots, \alpha_{i_l})$ units of input. Then we can define the short-run production possibilities of an industry with $N$ units during current period by a finite family of production activity vectors $\{a_i\}_{1 \leq i \leq N}$. Notice that, any vector $a_i$ from the collection of vectors $\{a_i\}_{1 \leq i \leq N}$ in $\mathbb{R}_+^{l+1}$ can be associated with a line segment

$$[0, a_i] = \{s_i a_i | s_i \in \mathbb{R}, 0 \leq s_i \leq 1\}.$$

Further with the assumption of $N \geq l + 1$, Hildenbrand defines the *short-run total production set* associated to the family $\{a_i\}_{1 \leq i \leq N}$ as the Minkowski sum

$$Y = \sum_{i=1}^{N} [0, a_i]$$

of line segments generated by production activities $\{a_i\}_{1 \leq i \leq N}$ and more explicitly, *short-run feasible industry production function* as the Zonotope

$$Y = \{y \in \mathbb{R}_+^{l+1} | y = \sum_{i=1}^{N} \phi_i a_i, 0 \leq \phi_i \leq 1\}.$$

Hildenbrand also defines his *short-run efficient industry production function* within the zonotope framework. Let's project the above-defined zonotope $Y$ on its first $l$ coordinates and denote this projection $D$, which reads,

$$D = \{u \in \mathbb{R}_+^l | \; \exists \, x \in \mathbb{R}_+ \; s.t. \; (u, x) \in Y\},$$

thus his production function $F \; : \; D \to \mathbb{R}_+$ follows,

$$F(u) = max\{x \in \mathbb{R}_+ | (u, x) \in Y\}.$$

This definition implies that given the level $u_1, \cdots, u_l$ of inputs for the industry, the maximum total output could be achieved by allocating, without any restrictions, the amounts $u_1, \cdots, u_l$ of inputs over the individual production units within the industry in one of the most efficient ways. However, the frontier associated with this production function does not provide any information on the actual technological set-up of the whole industry. This production function could not be the focal reference either from a positive or from a normative point of view (Hildenbrand 1981).

Within the Zonotope framework, Dosi et al. (2016) define the main diagonal of a Zonotope $Y$ as the diagonal joining the origin $O = (0, \cdots, 0) \in Y \subset \mathbb{R}^{l+1}$ with its opposite vertex in $Y$ and call this diagonal *production activity of the industry*, since it expresses both the amount of inputs employed and output produced by the industry. To be specific, this diagonal, denoted by $d_Y$, is simply the sum of individual production activities of the $N$ production units involved in the industry, i.e.

$$d_Y = (\beta_1, \cdots, \beta_l, \beta_{l+1}) = \left( \sum_{i=1}^{N} \alpha_{i_1}, \cdots, \sum_{i=1}^{N} \alpha_{i_l}, \sum_{i=1}^{N} \alpha_{i_{l+1}} \right) \in \mathbb{R}_+^{l+1} . \quad (2)$$

Obviously, if all firms in one industry were to use the same technique in a given year, all the vectors-firms would lie on the same line. This is the case where only one technology is adopted and all the firms within this industry are homogeneous. In this case, the associated Zonotope would degenerate to one with null volume, i.e. coinciding with the diagonal $d_Y$. On the other hand, the maximal heterogeneity case occurs when one industry involves some firms with almost zero inputs but sufficient output and others with a large quantity of inputs but little output. In such a case, the generated Zonotope almost becomes a parallelotope. Starting from this simple observation on these two extreme cases, it is possible to derive a rigorous measure for industry heterogeneity.

First, let $A_{i_1,\cdots,i_{l+1}}$ be the matrix whose rows are vectors $\{\alpha_{i_1}, \cdots, \alpha_{i_{l+1}}\}$ and $\Delta_{i_1,\cdots,i_{l+1}}$ its determinant. It is possible to compute the volume of the zonotope $Y$ in $\mathbb{R}^{l+1}$ employing the following formula:

$$Vol(Y) = \sum_{1 \le i_1 \le \cdots \le i_{l+1} \le N} |\Delta_{i_1,\cdots,i_{l+1}}| \quad (3)$$

where $|\Delta_{i_1,\cdots,i_{l+1}}|$ is the module of the determinant $\Delta_{i_1,\cdots,i_{l+1}}$. However, the value of $Vol(Y)$ depends both on the unit in which inputs and output are measured and on the number of firms. To normalize the measure, the volume of the Zonotope $Y$ generated by the production activities $\{a_i\}_{1 \le i \le N}$ is divided by the volume of the parallelotope with diagonal $d_Y = \sum_{i=1}^{N} a_i$. Notice that the parallelotope is the Zonotope with largest volume if the main diagonal is fixed. Such ratio is defined as:

$$G(Y) = \frac{Vol(Y)}{Vol(P_Y)} \quad (4)$$

where $P_Y$ denote the parallelotope with diagonal $d_Y$ and volume $Vol(P_Y)$. The normalized volume, G(Y), is named the *Gini volume*.[4]

Besides the heterogeneity measure, Dosi et al. (2016) also suggest that the angle formed by the industry production activity vector $d_Y$ with the space generated by all inputs expresses the industry productivity; and the tangent of this angle can be an appropriate measure.[5] To be specific, the measure of

[4] For a generalization of the measure, refer to Franciosi et al. (2020).

[5] Of course note that such intuition can be directly expanded to the level of the single vector-firm so as to derive an appropriate measure of productivity at the level of the firm.

productivity $P$ for a given industry including $N$ firms at current period, reads

$$P = tg\left(\Theta_{l+1}(d_Y)\right) = \frac{\sum_{i=1}^{N} \alpha_{i_{l+1}}}{||pr_{-(l+1)}\left(d_Y\right)||} \tag{5}$$

where for any vector $v = (x_1, \cdots, x_k) \in \mathbb{R}^k$ for $k = 2, 3, 4, \cdots$, projection map $pr_{-j}(\cdot)$ follows

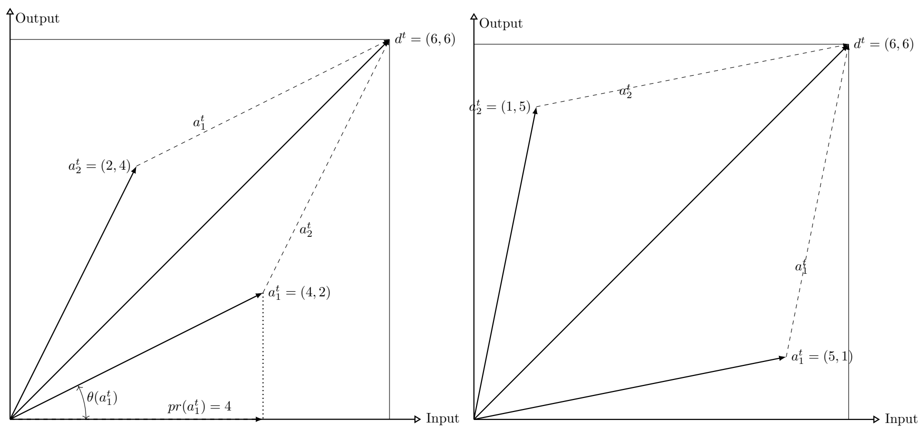
$$\begin{aligned} & pr_{-j}(v) : \mathbb{R}^k \to \mathbb{R}^{k-1} \\ & (x_1, \cdots, x_k) \mapsto (x_1, \cdots, x_{j-1}, x_{j+1}, \cdots, x_k), \end{aligned}$$

$\Theta_j(v)$ represents the angle formed by the vector $v$ and the space generated by all entities in vector $pr_{-j}(v)$ and $||v||$ represents the normal of the vector $v$. Furthermore, similarly, given the industry input vector $w = pr_{-(l+1)}(d_Y) \in \mathbb{R}^l_+$, we have $\Theta_i(w)$ and its tangent value to measure the relative intensity of input $i$ with respect to all the other inputs for $i = 1, \cdots, l$.

A few remarks are needed on the nature of the measure, its relation with similar techniques and the scope of applicability. Notice indeed, that the volume of Zonotope, as well as the angles, are measures and not estimates. Further, although the measure of productivity bears some similarity, mostly because of the non-parametric nature, with Data Envelopment Analysis in production (see, among the others, Farrell 1957; Charnes et al. 1978; Simar and Zelenyuk 2011) the two methods are clearly distinct, as in the Zonotope approach one is considering *all* firms, and not only those on the frontier. Although both approaches are data-driven and non-parametric in nature, the emphasis of DEA is, for the industry, to construct the efficient frontier by enveloping the data and, for individual firm, to proxy its efficiency (for an application to Stata, see Ji and Lee 2010; Badunenko and Mozharovskyi 2016). Empirically, the traditional deterministic approach faces some issues, which have been addressed by recent work. To be specific, Daraio and Simar (2007) deal with the sensitivity to measurement errors and outliers and Cazals et al. (2002); Daraio and Simar (2005) propose robust frontiers. There also exists a Stata command that allows to deal with such issues, see Belotti et al. (2013).

Finally, as also recalled above, notice that the Zonotope approach provides a general framework not necessarily limited to industry heterogeneity and productivity analysis. For example, Aruka (2017) argues that the Zonotope framework provides a different view of production set and also discloses new possibilities to model international trade. Indeed, in the latter field, he suggests a further generalization of the model so that it is possible to jointly assess more than 3 country and 3 commodities, thus abandoning the limiting special case of two countries-two commodities toy example. Or, referring to a rather different domain, the software application we are presenting here can be employed to compute measures of multivariate disparity similar to those outlined in Koshevoy and Mosler (1996, 1997).

Summing up, the proposed framework allows to investigate production and productivity dynamics of firms, while taking into consideration the existing




(a) Case 1 (b) Case 2

Figure 3: One-input, one-output examples of Zonotopes $Y = \sum_{i=1}^{2}[0, a_i^t] \subset \mathbb{R}^2$

heterogeneity across firms, which is high and persistent even within the same industry. The geometry based approach enables, on the one side to relax many of the existing assumptions currently present in the literature and which have been shown to often lack empirical support. On the other side, while it allows to provide a representation in the production space with several inputs and outputs, on a two or three dimensions, it is a way to provide a visual illustration of the trend of the industry and of the single firms.

To reinforce the last point, we take advantage of the simple one-input, one-output setting, as depicted in Figure 3a. There are two firms respectively represented by $a_1^t$ and $a_2^t$ as defined in (1). $a_1^t = (4, 2)$ indicates that firm 1 produces 2 units of outputs using 4 units of inputs. The industry production activity, as the sum of firm-vectors defined in equation (2), is represented by $d^t = (6, 6)$. In Figure 3b, we have identical industry production activity but with more firm heterogeneity. Comparing these two plots, the higher heterogeneity is consistent with the bigger parallelogram (the areas are equal to 12 and 24 respectively) which indicates the possibility of using volume of zonotope defined in (3) to measure the heterogeneity. However, as discussed, one still needs to get a more refined measure, by applying some normalization (as in equation (4)) to ensure that differences are not due, for instance, to different unit of measures, which might well happen in the real world. More in detail, the area is normalized by dividing it with the area of the corresponding square. This graphical illustration can also be employed to provide the intuition behind the productivity measures for the whole industry (defined in eq. (5)) and for the single firm (defined in eq. (6)) in $\mathbb{R}^2$. In this respect, figure 3a shows that the tangent value of $\Theta(a_1)$ is indeed the the ratio of firm 1's output over its input, i.e. labor productivity, which is one of the most popular productivity measures when taking only 1 input into account.

We conclude the section with table 1 which summarizes the key concept of the proposed zonotope method.

Table 1: Key concepts of the proposed method

| Key concepts | Details |
| --- | --- |
| Individual Production Activity $a_i$ | It is a vector defined in $\mathbb{R}_+^{l+1}$. Intuitively it represents the multiple-input, one-output production activity of one firm whose first $l$th elements are inputs while the $(l+1)$th is output |
| Zonotope $Y$ | It provides a description of the whole industry obtained combining together all firms. Formally, it is the Minkowski sum of line segments generated by firm production activities $\{a_i\}_{1 \leq i \leq N}$. |
| Gini $G(Y)$ | It is the normalized volume of the zonotope and it provides a measure for industry heterogeneity, the bigger the index, the larger the differences across firms. |
| Productivity $P = tg\left(\Theta_{l+1}(d_Y)\right)$ | In a two-inputs, one output case, the "steeper" the vector-firm (the vector-industry) the more productive the firm (the industry). More formally it is the tangent value of the angle formed by the vector and the input-hyperplane. |

# 3 The Zonotope Stata command

In this section we introduce the Stata command **zonotope**.[6]

## 3.1 Syntax

The syntax of the command to compute the zonotope is:

`zonotope` *varlist* `[if] [in] [, verbose]`

where the option `verbose`, when used, causes the program to print on screen all the computed quantities (see below). The zonotope command help is provided in Appendix A, and can be obtained from within Stata by entering `help zonotope`; instruction for installing the package under different Operating Systems are provided in Appendix B.

[6]Also note that in order to ease future developments, the software is available also at `https://github.com/zonotopes`

## 3.2 Description

The zonotope command takes as input a list of vector variables, where the last one is interpreted as the output and the others as the inputs of the relation we want to analyze. All the vector variables must have the same length. All the variables can be seen as columns of a single matrix *gen*, i.e., the matrix of generators, therefore its $i$-th row coincides with the $i$-th generator $a_i$, as defined in (1). The number of columns of this matrix represents the dimension where the zonotope lies, $(l+1)$. Notice that in the zonotope command, the generators are assumed to be non-negative, thus all entries of this matrix should be non-negative.

**Output and return value**

The zonotope command returns two vectors: *diagonal* and *tangents* and one matrix, named *gen* (it contains the generators actually used, an important information when the [if] and/or the [in] clauses are used). In addition, it returns a number of scalar values, such as

- `nrow`: the number of generators actually used (i.e., $N$);
- `ncol`: the number of variables (it coincides with $l+1$, if the program has been called with $l$ input variables and one output variable);
- `etMIN`: the elapsed time (expressed in minutes);

and the eight statistics `S1`, `S2`, ..., `S8` which are detailed below.

Finally, when the zonotope command is called with the *verbose* option, it shows all the returned variables (both vectors and scalars) on screen.

All the scalar variables returned by `zonotope` can be accessed using the `r()` command. As an example, the volume of the zonotope can be displayed using the command

```
.  display r(S1)
```

while the elapsed time using

```
.  display r(etMIN)
```

**Output vector: diagonal**

The output vector `diagonal` contains $d_Y$, the geometric diagonal of the zonotope which, according to (2), is defined as the sum of generator $a_i$ for $i = 1, \cdots, N$. To be specific,

$$d_Y = (\beta_1, \cdots, \beta_l, \beta_{l+1}) = \left( \sum_{i=1}^{N} \alpha_{i_1}, \cdots, \sum_{i=1}^{N} \alpha_{i_l}, \sum_{i=1}^{N} \alpha_{i_{l+1}} \right).$$

Clearly, it is an $(l+1)$-dimensional (row) vector. It can be easily displayed on screen (or reused) using `matrix list diagonal`.

**Output vector: tangents**
The output vector `r(tangents)` contains the tangent of the angle formed by each generator and the input space. Thus it is an $N$-dimensional (column) vector. It can be easily displayed on screen (or reused) using `matrix list tangents`.

**Output matrix: gen**
The output matrix `gen` contains the generators actually used, as filtered by in/if clauses. It can be easily displayed on screen (or reused) using `matrix list gen`.

We report below a brief description of the eight statistics mentioned above. Note that some of them have a clear economic interpretation, as they directly correspond to the key concepts of the proposed method, while others are necessary intermediate steps towards the measure of interest.

- **Statistic S1: Volume**

  Given $N$ generators $a_i \in \mathbb{R}_+^{l+1}$, we can generate one zonotope, again denoted as $Y$. Thus we can compute its volume as
  $$S1 \equiv Vol(Y)$$
  where $Vol(\cdot)$ follows (3). We report this volume as $S1$ which is printed on screen and also return it as an output variable (see above). S1 (Volume) provides a very preliminary idea of industry heterogeneity, but represents an intermediate step to compute the more refined measure of the Gini index.

- **Statistic S2: Diagonal's Norm**
  The norm of the diagonal $||d_Y||$ is computed as the square root of sum of the squares of all the components of the diagonal, i.e.
  $$S2 \equiv ||d_Y|| = \sqrt{\sum_{i=1}^{l+1} \beta_i^2}\ .$$
  The length of the diagonal of the zonotope represents the "size" of the industry, the longer this vector the bigger the industry is.

- **Statistic S3: Sum of Squared Norms of all the generators**
  As indicated by the name, we first compute, for each generator, the sum of the square of each of its components and then summing up over all generators. To be specific,
  $$S3 = \sum_{i=1}^{N} \left( \sum_{j=1}^{l+1} \alpha_{i_j}^2 \right)\ .$$

- **Statistic S4: Gini index**
According to (4), this Gini index is computed as the ratio between the volume of the zonotope and the product of the components of the diagonal. We re-write this industry heterogeneity measure as follows,

$$S4 \equiv G(Y) = \frac{Vol(Y)}{Vol(P_Y)} = \frac{S1}{\prod_{j=1}^{l+1} \beta_j}$$

where $P_Y$ denote the parallelotope with diagonal $d_Y$. The normalized volume of the zonotope, as already discussed, is a more refined measure of industry heterogeneity. The bigger this index, the more heterogeneous the industry.

- **Statistic S5: Tangent of angle formed by diagonal and input space**
Given the diagonal vector reported as $d_Y$, according to (5), we can further compute the tangent of angle formed by the diagonal and its input space. We report this industry productivity measure as $S5$ as follows,

$$S5 \equiv tg\left(\Theta_{l+1}(d_Y)\right) = \frac{\beta_{l+1}}{\sqrt{\sum\limits_{j=1}^{l} \beta_j^2}} .$$

Intuitively it provides our measure for industry productivity. In a two-inputs, one output setting the "steeper" the vector, the more productive is the industry.

- **Statistic S6: Cosine against output**
$S6$ reports the cosine of the complementary angle of $\Theta_{l+1}(d_Y)$ as

$$S6 = \frac{\beta_{l+1}}{\sqrt{\sum\limits_{j=1}^{l+1} \beta_j^2}} .$$

Due to the complementary angle relationship, also notice the connection between $S5$ and $S6$ as

$$S5 = \frac{S6}{\sqrt{1-(S6)^2}} .$$

- **Statistic S7: Cosine of diagonal projected on the input plane with x axis**
The angle formed by the $x$ axis and the projection of diagonal in the input plane measures relative intensity of the first input related to the other inputs. We report its cosine value as

$$S7 = \frac{\beta_1}{\sqrt{\sum\limits_{j=1}^{l} \beta_j^2}} .$$

- **Statistic S8: Volume against the cube of the norm of the diagonal** This last statistic is the ratio between the volume of the zonotope and the cube of the norm of the diagonal. To be specific,

$$S8 = \frac{Vol(Y)}{||d_Y||^3} = \frac{S1}{(S2)^3} .$$

  This definition of volume is restricted to the boundary edges of the cone of all possible vectors and as such it considers only the most diverse vectors. It attempts to measure the maximal diversity of the field, (hence it takes into account only the boundary vertices), by measuring how "wide" is the cone.

Other than such statistics, before quitting the command provides the elapsed time, expressed in minutes.

## 3.3 A working example of the zonotope command

In this section we provide a step-by-step working example that shows the use of the zonotope command with comments on the related output. The general framework, as before, is production analysis. The following section provides more examples, first, still in the field of productivity analysis, on a larger sample, and then considering other domains of applications for the zonotope command.

We start by loading into Stata the dataset and listing the first five rows

```
.  use empirical_firm_data.dta, clear
.  list in 1/5
```

| Year | NACE | Number∼s | FixedA∼s | Turnover | ID |
|---|---|---|---|---|---|
| 2006 | 0001 | 19 | 2969.52 | 29127.35 | 145725 |
| 2006 | 0001 | 9 | 920.53 | 2796.362 | 328840 |
| 2006 | 0001 | 22 | 631.447 | 4316.787 | 485189 |
| 2006 | 0001 | 35 | 2620.777 | 13958.58 | 227378 |
| 2006 | 0001 | 11 | 982.525 | 1362.247 | 491850 |

The first column indicates the year and the second reports the sector of main activity of the firm. Then, columns 3 to 5 report respectively, number of employees, fixed assets, and turnover. The last column prints a fake ID of the firm. In the interest of replicability of the findings produced with the new command, we are providing artificially generated data that display the same distributional properties as the "real" data employed in other works, such as Dosi et al. (2016, forthcoming). Each row in the dataset represents one firm in a specific year/industry. It is easy to see that all the first 5 firms in the dataset belong to sector "0001" in 2006.

As usual in the literature, number of employees and fixed assets are employed as proxies for labor and capital inputs, respectively, and turnover for output.

In this 2-inputs 1-output setting, the zonotope command allows to analyze the industry in $\mathbb{R}^3$. For example, focusing on a specific year and sector, one can write

```
.  zonotope NumberOfEmployees FixedAssets Turnover if NACE == "0001"
& Year == 2006, verbose
```

Notice that, although all the following results are actually printed in the terminal at the same time, we are going through them in a step-by-step fashion so as to provide specific comments to the results that are relevant to the investigation.

---

ZONOTOPE LIBRARY VER 1.3

---

INPUT: SET OF GENERATORS
N. of dimensions (including the output): 3
N. of generators: 160

---

This first part of the output, after informing about the version of the package, displays the number of dimensions of the current analysis, three in the example, and the number of generators that in this case are firms.

---

Computation started (it can take a while) ...

---

OUTPUT VECTORS
Diagonal (a row vector):
5531 459015 2.84636e+06

---

The second part of the output reports the production activity of the whole industry, the diagonal of the zonotope, as defined by equation (2) that is the result of the aggregation of the 160 generators-firms. It is easy to see that, the total number of employees of this industry is 5531 while the total fixed assets and turnover of this industry are equal to 459,015 and 2,846,360 thousand Euros respectively.

---

Tangent between each input generator and the input space:
9.80857
3.03763
6.8322
5.32565
1.38639
(...)

———————————————————————————————————

For the sake of the exposition we have reported the tangent only for the first five firms of this example. As discussed in details in section 4.2, we extend the definition of industry productivity, eq. (5), to individual firm productivity, as proxied by the angle that the vector-firm forms with the input plane. For example, firm "145725" produces output 29127.35 thousand Euro from 19 employees and 2969.52 thousand Euro fixed asset. According to equation (6), productivity of firm "145725", as proxied by the tangent of the above mentioned angle, is

$$\frac{29127.35}{\sqrt{19^2 + 2969.52^2}} = 9.80857 \, .$$

Notice that a higher value of the tangent of the angle suggests a higher productivity as the firm-vector is steeper: the angle that the vector forms with input plane is larger. Figure 3a provides a simple illustration of this, in a one-input, one-output setting.

As for many other indicators of efficiency that employ multiple inputs, the value of the measure alone is not of great use, but it is much relevant instead to perform comparison across firms within the same industry or inter-temporal comparison of the same unit over time. This is indeed the sort of comparison that is proposed in Table 2 below.

———————————————————————————————————

OUTPUT STATISTICS
S1: Total volume: 1.46438e+15
S2: Diagonal norm: 2.88313e+06
S3: Sum of squared norms: 2.9266e+11
S4: Gini index: 0.202645
S5: Tangent of angle btw. diagonal and the input plane: 6.20056
S6: Cosine against output: 0.987243
S7: Cosine of proj. of diagonal on input plane with x axis: 0.0120488
S8: Volume against the cube of the norm of the diagonal: 6.11027e-05
———————————————————————————————————

The zonotope command then produces the eight statistics already described at length in the previous section. In this working example we only recall some of them. For example, `S2` reports the length of the diagonal which measures the size of the industry taking two inputs and one output into account. `S4` reports the Gini index which measures the heterogeneity of the sector while `S5` reports industry productivity, as defined in equation (5). Let us emphasize that we strongly suggest to employ the measure of heterogeneity of a sector (or, in a different context, of the relevant aggregate) to make cross-sectional or intertemporal comparison. This is how the measures of heterogeneity are employed in Dosi et al. (2016, forthcoming). The same argument applies to `S5` and `tangents`, i.e., the tangent value of the angle between the diagonal and the input plane. For example, it is difficult to conclude that one industry in one specific country is highly productive or not only based on the magnitude of `S5`. But if we have productivity levels `S5` for the same industry in two different countries, then it is possible to compare which country is more productive in a given industry. Or if we have productivity level `S5` for one industry in one specific country over time, it is then possible to determine whether that country has recorded a productivity increase in that industry. In section 4.1, we provide a more detailed example regarding the dynamics of heterogeneity and productivity for a few selected industries over time. `S7` reports the cosine value of the angle formed by the projection of diagonal on input plane, i.e., industry input vector $(5531, 459015)$ with the x axis, i.e. the axis corresponds to the number of employees. The value 0.0120488 indicates that this angle is almost $\pi/2$, the industry adopts much more fixed asset compared with employees. Notice that the value of this angle is highly influenced by the choice of the unit of measure, for a discussion of this issue, see Dosi et al. (2016, forthcoming).

---

Elapsed time (MIN):
0.001383

---

Last, we have the computation time reported. As it will be explained at greater detail in Section 5, building the zonotope of a set of generators has an exponential complexity (i.e., it is very time consuming, especially in high dimension and when the number of generators in high).

Finally it is possible to display (or reuse) the computed results as follows:

---

display r(S1) (for the volume)
display r(S2) (for the norm of the diagonal)
...
display r(S8) (volume against the cube of the norm of the diagonal)

matrix list diagonal (for the vector containing the diagonal)
matrix list tangents (for the vector with the tangents)
matrix list gen (for the actually used generators, after if/in)

---

# 4 Empirical analysis

In this section, using the zonotope command, we provide three different more advanced examples related to performing economic investigations. First, based on firm-level data, we compute the heterogeneity and productivity levels of industries as suggested by Dosi et al. (2016). Second, firm-level measures of productivity are computed resorting to zonotope command. Third, based on the income and expenditure data of Namibia, we show that the zonotope command also enables to compute multivariate Gini coefficient as a measure for inequality as suggested by Mosler (1994); Koshevoy and Mosler (1997).

## 4.1 Assessing Industry Heterogeneity

We now conduct some empirical investigations on the industry heterogeneity and productivity and we do this by employing larger sets of data which are meant to replicate the standard situation faced during empirical analysis on firm level data, where firms are grouped according to their sector of main activity, at different level of aggregation.

To allow for the replication of all the analysis we employ the same set of data introduced in the previous section. For each firm in the selected industries, number of employees and fixed assets are chosen to proxy inputs and turnover for output. All of these values, except for the number of employees, are assumed to be in thousands euro and deflated at 4-digit NACE level with year 2010 as benchmark year to perform inter-temporal comparison. Note that we refer to the 4-digit NACE sector simply as “0001”, “0002”, etc. When it is necessary to aggregate 4-digit sectors into 3-digit sector, we use “0091”, “0092” and “009” as in figures 1 and 2.

### 4.1.1 The two-inputs one-output case

Traditionally, the most standard setting to investigate production is that which assumes two inputs and one output production activity; number of employees and fixed assets are chosen to be the proxies for inputs and turnover value for the output. For each industry in one specific year, the normalized zonotope volume and the tangent value of the angle formed by industry production vector and its input plane are easily computed as S4 and S5 that can be used to measure the industry heterogeneity and productivity level for one specific industry at a given point in time. Columns (2) and (3) of Table 2 report these two results for selected industries in 2006 while column (1) reports the number of firms within that industry-year cell. To get these three variables for a specific industry in a specific year, for example sector 0001 in 2006, we run the following commands

```
.  use empirical_firm_data.dta, clear
.  zonotope NumberOfEmployees FixedAssets Turnover if Year == 2006
& NACE == "0001", verbose
```

from the available results we select

———————————————————————————————————

ZONOTOPE LIBRARY VER 1.3

———————————————————————————————————

N. of generators: 160

———————————————————————————————

S4: Gini index: 0.202645
S5: Tangent of angle btw. diagonal and the input plane: 6.20056

———————————————————————————————

Table 2: Computation Results for Gini Coefficient and Productivity among Selected Sectors and Years - $\mathbb{R}^3$ Case

| | **Year 2006** | | | **Year 2009** | | | **Year 2012** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Sectors** | **Obs** | **Gini** | **tg** | **Obs** | **Gini** | **tg** | **Obs** | **Gini** | **tg** |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| 0001 | 160 | 0.203 | 6.201 | 222 | 0.233 | 4.239 | 346 | 0.264 | 4.483 |
| 0002 | 115 | 0.126 | 6.617 | 134 | 0.119 | 6.780 | 162 | 0.138 | 9.292 |
| 0003 | 14 | 0.016 | 1.561 | 31 | 0.020 | 1.803 | 70 | 0.062 | 1.385 |
| 0004 | 386 | 0.153 | 3.014 | 431 | 0.190 | 2.077 | 603 | 0.200 | 2.937 |
| 0005 | 57 | 0.090 | 5.744 | 69 | 0.087 | 3.116 | 142 | 0.114 | 4.590 |
| 0006 | 142 | 0.155 | 4.078 | 168 | 0.170 | 2.660 | 212 | 0.254 | 5.859 |
| 0007 | 43 | 0.098 | 0.864 | 52 | 0.112 | 0.892 | 68 | 0.123 | 1.265 |
| 0008 | 47 | 0.091 | 3.402 | 53 | 0.101 | 2.292 | 69 | 0.122 | 3.107 |
| 0009 | 38 | 0.073 | 1.495 | 41 | 0.088 | 1.060 | 47 | 0.080 | 2.204 |
| 0010 | 18 | 0.084 | 1.276 | 26 | 0.092 | 1.048 | 40 | 0.201 | 1.995 |
| 0011 | 65 | 0.103 | 3.834 | 85 | 0.172 | 3.249 | 126 | 0.168 | 3.487 |
| 0012 | 177 | 0.204 | 3.102 | 189 | 0.155 | 2.153 | 259 | 0.164 | 3.153 |
| 0013 | 54 | 0.042 | 2.372 | 58 | 0.063 | 1.539 | 82 | 0.076 | 1.796 |
| 0014 | 225 | 0.137 | 3.376 | 245 | 0.125 | 2.218 | 368 | 0.126 | 2.755 |
| 0015 | 28 | 0.001 | 0.899 | 33 | 0.001 | 0.725 | 45 | 0.001 | 0.751 |
| 0016 | 45 | 0.026 | 1.288 | 44 | 0.022 | 0.912 | 62 | 0.042 | 0.875 |
| 0017 | 86 | 0.072 | 3.713 | 89 | 0.089 | 1.486 | 121 | 0.094 | 2.292 |

The first observation is that the chosen *normalization* strategy for the volume of the zonotope seems to be effective, as there is no apparent relation between the number of firms-generators and the Gini coefficients. For example, in 2006, there are 386 firms in the 4-digit sector 0004 while 160 firms in sector 0001. However given a larger number of firms, we don't expect that Gini coefficient, i.e. the normalized volume of zonotope, of sector 0004 to be necessarily bigger than that of sector 0001. Indeed, based on our data sample, the Gini coefficient of 4-digit sector 0001 is larger, 0.203 as compared to 0.153 of sector 0004.

After taking into account the effect due to the number of firms, we notice that among different industries, the heterogeneity levels are different. For example, in 2006 the Gini coefficients vary from 0.001 for 4-digit sector 0015 to 0.204 for 4-digit sector 0012. Similarly, as indicated in column (3), in the same year, the productivity levels among different sectors are also different.

From column (4) to (6) and from column (7) to (9), we report similar results in years 2009 and 2012 respectively. This allows us to explore the dynamic of industry heterogeneity and productivity over time. We notice that most of the selected sectors share an upward trend in their heterogeneity levels. For example, heterogeneity level of 4-digit sector 0001 increases from 0.203 in 2006 to 0.233 in 2009, and to 0.264 in 2012. As to the productivity we report, for most of industries, a decrease from 2006 to 2009 and an increase from 2009 to 2012.

Finally notice that the same analysis can be performed in the case of 4-Dimensions, for instance in the 3 inputs and 1 output case[7]. In the interest of space, results are not reported here but they can be obtained with the following commands

```
. use empirical_firm_data.dta, clear
. zonotope NumberOfEmployees FixedAssets MaterialCost Turnover if
Year == 2006 & NACE == "0001", verbose
```

## 4.2 A geometric approach to firm-level productivity

As indicated in equation (5), Dosi et al. (2016) propose the tangent of the angle formed by the industry production activity vector $d_Y$ with the space generated by all inputs as a measure for industry productivity. The approach can be easily applied also to individual firm. Assume firm $i$'s production activity follows equation (1), then its productivity level can be measured by tangent of $\Theta_{l+1}(a_i)$, i.e. the angle formed by the firm production activity vector $a_i$ with the space generated by all inputs. To be specific, one possible measure for productivity of firm $i$, denoted by $p_i$, follows,

$$p_i = tg\left(\Theta_{l+1}(a_i)\right) = \frac{\alpha_{i_{l+1}}}{||pr_{-(l+1)}\left(a_i\right)||} \tag{6}$$

for $i = 1, \cdots, N$.

The zonotope command computes this productivity for each firm and returns all of them as the vector *tangents*. Based on the same firm-level data employed above, we compute the productivities of twelve selected firms from industry 4-digit sector 0010 in 2006 and report them in the column "Year 2006" of table 3. Also this newer measure is consistent with the stylized fact from the previous literature in that we still observe relevant heterogeneous productivity at firm-level. To compute the productivity (6) for the 12 firms reporting data in all years in sector 0010 in 2006[8], we run the following commands

```
. use empirical_firm_data.dta, clear
. destring ID, replace
. egen selected = anymatch(ID), values(40029 84229 92290 96040 117252
```

[7] In this case, material cost is chosen to be the proxy for the third input.

[8] For illustrative purpose the code only displays results for year 2006. To generate the columns for other years it is enough to change the "if" condition. The do file that is provided with the package generates the complete Table 3 as reported in the text.

```
140293 162695 214062 214655 227040 233379 257599)
.  keep if selected
.  sort ID NACE Year
.  zonotope NumberOfEmployees FixedAssets Turnover if Year == 2006
& NACE == "0010", verbose
```

among the available results we focus on that referring to firm productivity

---

Tangent between each input generator and the input space:
1.34676
6.91531
0.399336
5.51919
0.851644
5.66018
11.7067
23.4575
13.3015
4.76769
0.46963
0.962496

---

We then compute productivity of these firms, if they stay in the industry, in 2009 and 2012 and report the results respectively in the columns "Year 2009" and "Year 2012" of the same table. Note that firm-level productivity, again in accordance with previous findings, display persistence, hence the large within-industry dispersion does not vanish over time.

Table 3: Productivity of Selected Firms from 4-digit sector 0010 - $\mathbb{R}^3$ Case

| **Firm ID** | **Year 2006** | **Year 2009** | **Year 2012** |
|---|---|---|---|
| 1 | 1.347 | 2.470 | 4.353 |
| 2 | 6.915 | 6.704 | 13.996 |
| 3 | 0.399 | 0.178 | 0.292 |
| 4 | 5.519 | 4.470 | 8.272 |
| 5 | 0.852 | 0.603 | 0.997 |
| 6 | 5.660 | 3.336 | 5.394 |
| 7 | 11.707 | 14.092 | 14.949 |
| 8 | 23.457 | 10.902 | 4.380 |
| 9 | 13.302 | 19.282 | 3.519 |
| 10 | 4.768 | 5.417 | 11.143 |
| 11 | 0.470 | 0.622 | 0.762 |
| 12 | 0.962 | 1.090 | 1.393 |

### 4.3 An application to Inequality

The zonotope command is not only limited to firm-level data to compute industry heterogeneity and productivity levels at industry and firm level. In this section, we perform another experiment to show that the command is also helpful for the computation of the multivariate Gini coefficient which measures inequality among a group of $N$ units with $l > 1$ attributes.

#### 4.3.1 Introduction to multivariate Gini coefficient

Following Mosler (1994); Koshevoy and Mosler (1996) we focus on two multivariate Gini coefficients. To do this, we denote by $\omega_i^j$ the quantity of attribute $j$ owned by unit $i$ for $i = 1, \cdots, N$ and $j = 1, \cdots, l$. Thus we have vectors

$$w_i = (\omega_i^1, \cdots, \omega_i^j, \cdots, \omega_i^l) \in \mathbb{R}_+^l$$

for $i = 1, \cdots, N$. Correspondingly,

$$\tilde{\omega}_i^j = \frac{\omega_i^j}{\sum_{i=1}^N \omega_i^j}$$

indicates the share of attribute $j$ out of the total amount owned by unit $i$ and vectors $\tilde{w}_i$ follows as

$$\tilde{w}_i = (\tilde{\omega}_i^1, \cdots, \tilde{\omega}_i^j, \cdots, \tilde{\omega}_i^l) = \left( \frac{\omega_i^1}{\sum_{i=1}^N \omega_i^1}, \cdots, \frac{\omega_i^j}{\sum_{i=1}^N \omega_i^j}, \cdots \frac{\omega_i^l}{\sum_{i=1}^N \omega_i^l} \right) \in \mathbb{R}_+^l$$

for $i = 1, \cdots, N$.

Mosler (1994) introduces a multivariate Gini index by using the volume of the Lorenz zonotope (LZ) as the Minkowski sum

$$\text{LZ} = \sum_{i=1}^{N} [0, (\frac{1}{N}, \tilde{w}_i)] \tag{7}$$

of line segments generated by $\{(\frac{1}{N}, \tilde{w}_i)\}_{1 \leq i \leq N}$. Obviously, the zonotope command provides this volume, i.e. $Vol(\text{LZ})$. Note however that Koshevoy and Mosler (1996) already point out that $Vol(\text{LZ})$ would be zero when two attributes are similarly distributed among $N$ units or one attribute is equally distributed among all units[9]. Hence, Koshevoy and Mosler (1997) suggest that instead of using the volume of LZ, one should employ the volume of the lift zonoid "expanded" by a $l$-dimensional cube in $\mathbb{R}^{l+1}$. Indeed, they propose the multivariate volume-Gini index as,

$$R_v = \frac{1}{2^l - 1} \sum_{s=1}^{l} \left[ \sum_{1 \leq j_1 \leq \cdots \leq j_s \leq l} Vol(Z^{j_1, \cdots, j_s}) \right] \tag{8}$$

[9]Let's construct one $N \times (l+1)$ matrix where row $i$ is $(\frac{1}{N}, \tilde{w}_i)$. Indeed, if the rank of this matrix is smaller than $(l+1)$, we would have the volume of the zonotope equal to zero. Besides the two possibilities we highlight in the text, note that if we have $N < l+1$, we also have that the rank is smaller than $(l+1)$. However, empirically, we generally have $N$ big enough to allow row rank is bigger than $(l+1)$.

where zonotopes are defined as the Minkowski sum

$$Z^{j_1,\cdots,j_s} = \sum_{i=1}^{N} [0, (\frac{1}{N}, \tilde{w}_i^{j_1,\cdots,j_s})]$$

of line segments generated by vectors $\{(\frac{1}{N}, \tilde{w}_i^{j_1,\cdots,j_s})\}_{1 \le i \le N}$ and $\tilde{w}_i^{j_1,\cdots,j_s} \in \mathbb{R}_+^s$ is obtained from the $j_1, \cdots, j_s$ components of vector $\tilde{w}_i$.

To provide the intuition of why it is necessary to resort to this further measure, let us refer to the toy example of appendix C and focus on $x$ and $z$, representing two attributes among the 10 units. Using the zonotope command, we can easily compute the volume of the corresponding LZ defined in equation (7) as 0.053. Now assume we have a third attribute that still coincide with $z$. We would then have volume of LZ equal to zero. On the other hand, multivariate volume-Gini index defined in equation (8) is able to provide a result 0.144 also when there are identically distributed attributes.

Notice that to get this volume-Gini index, we need to compute volumes of different $(2^l - 1)$ zonotopes. And for each computation, the zonotope command is employed. If we construct a sample made of normalized data, i.e. a $N \times (l+1)$ matrix with each row as $(\frac{1}{N}, \tilde{w}_i)$ where every column sums up to 1, to get each $Vol(Z^{j_1,\cdots,j_s})$ the zonotope command computes corresponding columns of data sample and report the result in S1. However, recall that the zonotope command also reports normalized volume of zonotope in S4 and this facilitates the process.

#### 4.3.2 Empirical application of multivariate Gini coefficient

We present one empirical exercise based on actual data from Namibia household Income and Expenditure Survey 2009/2010(The Namibia Statistics Agency 2013)[10]. This dataset includes detailed information of income and expenditure of 9,643 households in Namibia from 2009 to 2010. We compute, for each household, the income per capita (IPC) and expenditure per capita (EPC) based on its income, expenditure, and number of persons in the household[11]. Notice, in particular, that the survey reports only the main source of income for each household. As a result, we have two attributes of each household, IPC and EPC, either of which can be employed for computing the traditional univariate Gini coefficient.

In many studies, univariate Gini coefficient based on IPC is adopted as the measure of inequality. We report it in the end of column (1) of table 4. Correspondingly, the univariate Gini coefficient based on EPC is reported at the bottom of column (2). The univariate Gini coefficient indicates higher inequality level based on expenditure than income with 0.78 compared with 0.611. As we discuss above, these two attributes can be taken into account simultaneously by

[10] We downloaded the data file `NAM_2009_HIES_v01_M_v01_A_PUF_STATA8.zip` on June 20, 2019 with MD5 value 90461521c1059fa1df3bf97e6b2dc797 from `https://nsa.org.na/microdata1/index.php/catalog/6/study-description`.

[11] According to the suggestion from the original dataset, a person beyond (including) 16 years old contributes 1 to the number, a person between 6 to 15 contributes 0.75, and a person younger than 5 contributes 0.5.

computing the volume of the correspondingly LZ and multivariate volume-Gini index. We report them at the bottom of columns (3) and (4) respectively. Notice that, although the values of these two multivariate Gini indices are smaller than the univariate one, it is not possible to compare the values of univariate and multivariate Gini indices since they are constructed in different ways.[12] On the other hand, it is meaningful to compare the index, based on same definition, cross section and/or over time. And we continue to conduct one cross-section analysis as follows.

Table 4: Various Gini Coefficients Among Households with Different Income Sources in Namibia from 2009 to 2010

| | **Univariate Gini** | | **Multivariate Gini** | | |
|---|---|---|---|---|---|
| **Income Source** | **IPC** | **EPC** | $Vol(\mathbf{LZ})$ | $R_v$ | **Obs** |
| | (1) | (2) | (3) | (4) | (5) |
| Salaries / wages | 0.558 | 0.700 | 0.155 | 0.471 | 4,953 |
| Farming | 0.656 | 0.866 | 0.206 | 0.576 | 2,014 |
| Business activities | 0.651 | 0.799 | 0.213 | 0.554 | 785 |
| Employment pension | 0.581 | 0.777 | 0.253 | 0.537 | 121 |
| Cash remittances | 0.465 | 0.645 | 0.127 | 0.413 | 282 |
| State old pension | 0.422 | 0.619 | 0.135 | 0.392 | 986 |
| Other, specify | 0.666 | 0.864 | 0.288 | 0.606 | 502 |
| Total | 0.611 | 0.780 | 0.190 | 0.527 | 9,643 |

Households are classified into seven groups according to their main source of income. We report the univariate and multivariate Gini coefficients respectively between column (1) and (4) in the same table, and the numbers of households per group in column (5). Notice first, that groups with different income source display different inequality levels. According to univariate Gini coefficient based on IPC, the least unequal group is the one relying on state old pension. This seems reasonable since state old pensions, compared to other sources, are expected to be more equally distributed among receivers. The second observation is related to the potential bias associated to $Vol$(LZ) as measure of inequality, as it is sometimes not consistent with the evidence stemming from the univariate Gini coefficients. For example, the group "Business activities" seems more unequal compared to "Employment pension" according to the two univariate Gini coefficient and volume-Gini index $R_v$ while $Vol$(LZ) indicates the opposite. This clearly exemplifies one possible drawback of $Vol$(LZ) as a measure of inequality. When IPC and EPC are more correlated, the value of $Vol$(LZ) turns to be smaller[13]. Thus, it is not impossible that two highly correlated attributes, both

[12]Both volumes of the correspondingly LZ and multivariate volume-Gini index degenerate to the univariate Gini coefficient when $l = 1$. However, when $l > 1$, they are different from the univariate Gini coefficient.

[13]The extreme case would be when two attributes are identically distributed among all units

pointing to high-level inequality according to univariate Gini coefficient, would point to low inequality when based on $Vol(\mathrm{LZ})$. Intuitively, the measure of the volume, $Vol(\mathrm{LZ})$, becomes smaller due to the correlation between such two attributes. In this case, although the two univariate Gini coefficients of group "Business activities" are bigger than those "Employment pension", since the correlation between IPC and EPC is bigger for the former (0.814 versus 0.652), $Vol(\mathrm{LZ})$ of "Business activities" becomes smaller than "Employment pension" group. The volume-Gini index is considered to be more robust than $Vol(\mathrm{LZ})$ not only when there are identically distributed attributes but also when there are highly correlated attributes.

We provide the code to generate the fourth row of Table 4. First of all, we generate the relevant variables, i.e. IPC and EPC from raw data following below codes.

```
. use 7a0_NHIES_2009_2010_Household_dataset.dta, clear
. keep if cMainInc == 5
. gen incomepercapita = hIncComp/ConFac2
. gen exppercapita = hF1ExpSm/ConFac2
. egen total = total(incomepercapita), by(cMainInc)
. replace incomepercapita = incomepercapita/total
. drop total
. egen total = total(exppercapita), by(cMainInc)
. replace exppercapita = exppercapita/total
. drop total
. egen NumOfObs = count(cMainInc), by(cMainInc)
. gen Frequency = 1/NumOfObs
```

Then we compute the univarite Gini coefficient based on IPC and EPC respectively using the Stata command `inequal7`.

```
. matrix Result = J(8,6,0)
. * compute univarite Gini based on income per capita
. inequal7 incomepercapita, returnscalars
. matrix Result[4, 2] = r(gini)
. * compute univarite Gini based on expenditure per capita
. inequal7 exppercapita, returnscalars
. matrix Result[4, 3] = r(gini)
```

In the end, we compute $Vol(\mathrm{LZ})$ and volume-Gini Index $R_v$ based on IPC and EPC where we need the package to compute the volume of different zono-

and $Vol(\mathrm{LZ})$ goes to zero.

topes.

```
.  tuples incomepercapita exppercapita, di
.  matrix tem = J(‘ntuples’,1,0)
.  forvalues i = 1/‘ntuples’ {
.  di ‘i’
.  zonotope Frequency ‘var’‘tuple‘i”
.  if ‘i’ == ‘ntuples’ {
.  matrix Result[4, 4] = r(S1)
.  matrix Result[4, 6] = r(nrow)
.  }
.  matrix tem[‘i’, 1] = r(S1)/(2∧2 - 1)
.  }
.  matsum(tem), col(ColsumTem)
.  matrix Result[4, 5] = ColsumTem
.  frmttable, statmat(Result) sdec(0,3,3,3,3,0) title("Gini Namibia
Case")
```

## 5 Analysis of Zonotope Computing Time

Computing the volume of the zonotope given the list of its generators can potentially be much time consuming, since its computational complexity is $\mathcal{O}(N^l)$, where $N$ is the number of generators and $(l+1)$ is the dimension of each generator (i.e., its length). Thus this algorithm falls within the category of exponential algorithm, i.e., its time does not scale in a polynomial way, but instead in an exponential way. In particular, when $N$ is kept constant, it is clearly exponential in the number of dimensions $(l+1)$.

Table 5 provides the elapsed times for a varying number of variables (from 2 to 6) and a fixed number of generators ($N = 200$). On the contrary, Table 6 provides the elapsed times for a varying number of generators (from 50 to 250) and a fixed number of variables ($(l+1) = 6$). We have run the experiments on an Intel CPU i7 4 cores on Windows 7 Operating System, with 32 Gb of RAM, using Stata/IC. The content of Tables 5 and 6 are charted on Figures 4 and 5, respectively. From such figures it is clear how the time complexity of the zonotope is exponential. To empirically prove that the time complexity is $N^l$, we have plotted in Figure 6 the fifth root of the elapsed times, when the number of variables $(l+1)$ is equal to 6 (the same setting of the second experiment). The fact that the graph is a perfect straight line proves that the dependency is a power of 5, as expected.

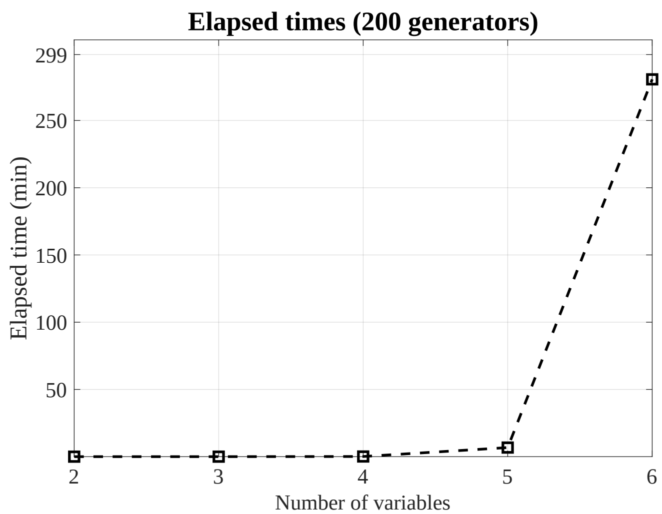



Figure 4: Elapsed time as a function of the number of variables, for 200 generators.

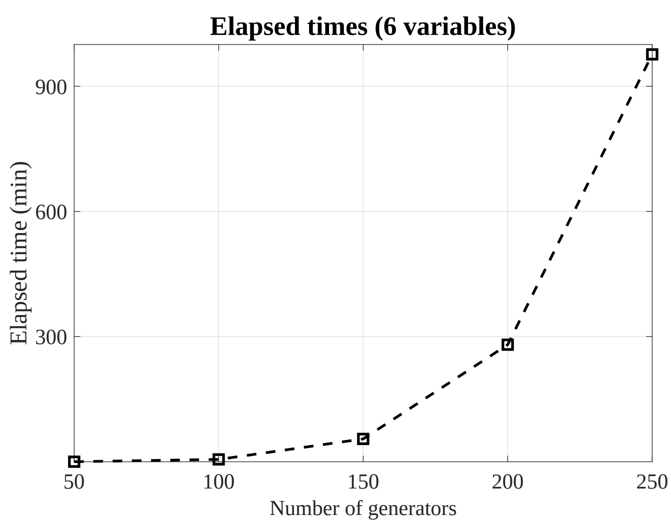



Figure 5: Elapsed time as a function of the number of generators, for 6 variables.

Table 5: Elapsed times for varying numbers of variables and 200 generators

| No. of vars | No. of generators | Elapsed time |
|---|---|---|
| 2 | 200 | 0.00017 min |
| 3 | 200 | 0.00157 min |
| 4 | 200 | 0.11477 min |
| 5 | 200 | 6.75457 min |
| 6 | 200 | 280.61160 min |

Table 6: Elapsed times for varying numbers of generators and 6 variables

| No. of vars | No. of generators | Elapsed time |
|---|---|---|
| 6 | 50 | 0.126 min |
| 6 | 100 | 5.648 min |
| 6 | 150 | 54.736 min |
| 6 | 200 | 4h and 40.612 min |
| 6 | 250 | 16h and 16.389 min |

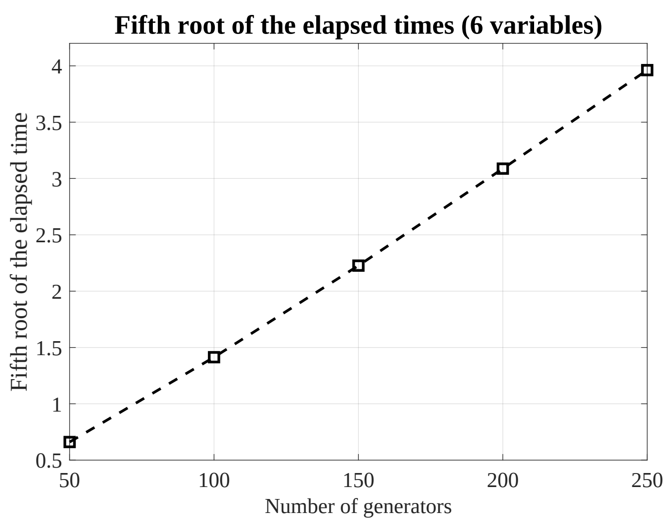



Figure 6: The fifth root of the elapsed time as a function of the number of variables, for 6 variables. The fact that the graph is a perfect straight line empirically proves that the computational complexity is $N^l$, since $(l+1) = 6$ in our case ($N = 200$, of course).

## 6 Conclusions and Future Work

The recent and increasing availability of disaggregated economic data contributed to challenge many of the standard assumptions from the theory, yet, at the same time, there is still urgent need for adequate tools to deal with the richness and complexity emerging from empirical observations.

In this respect the Zonotope package provides a rigorous way to perform empirical analysis taking advantage, and not neglecting some of these emerging properties. In this work we have proposed an application to production analysis in which, thanks to the proposed methodology, it is possible to relax most of standard assumptions that do not find support in the data. Firms, and economic agents more in general, are much different from each other under many respects: size, productivity, propensity to innovate and export, etc. Such differences do not vanish over time: selection, if at work, takes long to exerts its effects. Even focusing on firms within the same narrowly defined industrial sector does not help in reducing heterogeneity. In this context, the Zonotope package enables to assess the level of intra-industry heterogeneity and to measure the level of productivity, and its variation over time, without imposing strong assumptions on the actual observations. Finally, notice that the software package we are introducing is ready for applications in other domains, as the zonotope framework itself is already employed in other fields, for example, still within economic analysis, to assess inequality.

# A The zonotope help

**Title**
Welcome to the help for the **zonotope** command.

**Syntax**

**zonotope** *varlist* [if] [in] [, **v**erbose]

**Description**

This function builds the zonotope from a list of (l+1) vector variables, where the first l are the input ones and the last one is the output variable.

All the variables must have the same length.
The r-th value of each variables constitute the r-th 'generator', i.e., the r-th vector of the observations.

The command computes the volume and other quantities of the zonotope associated to this set of r generators in an (l+1) dimensional space.

The number of variables must be greater or equal to 2.

The function returns the volume of the zonotope in the scalar variable r(S1), plus additional statistics in r(S2),...,r(S8).

It also returns the elapsed time, in minutes, in the r(etMIN) variable, the vector with the diagonal, the vector with the tangents of the angles formed by each generator and the input space, and the matrix with the actually used generators (the latter is useful when if/in conditions have been used).

To diplay these arrays, use:

```
matrix list diagonal
matrix list tangent
matrix list gen
```

Finally, the zonotope command prints on screen all the computed quantities, when it is called with the 'verbose' option (only the matrix with the generators is not automatically displayed, since it would occupy too many lines on screen, in general).

**Remarks**

The zonotope command is an .ado program that calls a Stata plugin written using the C++ programming language.

The following GitHub repository contains the Stata command source code:

https://github.com/zonotopes/zono_stata

while the C++ source to generate the plugin can be found here:

https://github.com/zonotopes/zono_cpp

Further information about applications of the zonotope command to Economics can be found here:

G. Dosi, L. Marengo, M. Grazzi, and S. Settepanella (2016), "Production Theory: Accounting for Firm Heterogeneity and Technical Change". The Journal of Industrial Economics LXIV: 875-907.

**Examples**

. **run "`c(sysdir_plus)'/z/zonotope_demo"**

. **run "`c(sysdir_plus)'/z/zonotope_demo2"**

# B Getting started and plugin generation

To use the zonotope command, normally running the following Stata commands is enough:

```
. net cd http://www.iet.unipi.it/m.cococcioni/zono_stata_1p3
```

Then run:

```
. net install zonotope, replace force
```

To run the demo, execute the following do file:

```
. run "`c(sysdir_plus)'/z/zonotope_demo"
```

If the demo runs correctly, you are done.

However, it may fails, especially if you are using an operating system for which the plugin has not been generated yet.

In such case, the plugin must be generated from scratch, as explained in the next subsection.

## How to generate the zonotope plugin

First of all you need to download the C++ source code from the following git repository:

https://github.com/zonotopes/zono_cpp

It contains everything is needed to generate the Stata plugin, for most of the platforms (see below).

However, in order to generate the plugin, the following things are required:

- a C++ compiler
- the CMAKE utility
- (optional) the GIT utility (it is useful download the tree of the whole project or specific sub-projects)

The CMAKE utility is required in order to compile the plugin, as a shared library. Such shared library, named zonotope2.plugin or zonotope3.plugin (depending on the Stata version ) is used by the ado file associated with the zonotope command.

The latter file is named zonotope.ado and, together with demo files, additional datasets and help file can be download from:

https://github.com/zonotopes/zono_stata

The plugin generation has been tested on Windows 10 64bit using Visual Studio 15, on Linux Ubuntu 18.04 64bit using GCC and on MacOs 10.12.3 (16D32), 64 bit, using the MacOS Sierra operating system, with LLVM C++ compiler.

In case of troubles, please open an issue on the associated GitHub repository. We will do our best to help you.

Added on Aug 2, 2021

## C A simple toy example

As an example, by entering the following commands:

```
.  use http://www.stata-press.com/data/r11/test.dta
.  list
```

we will have the following dataset loaded into Stata memory:

| x | y | z |
|---|---|---|
| 1 | 10 | 2 |
| 2 | 9 | 4 |
| 3 | 8 | 3 |
| 4 | 7 | 5 |
| 5 | 6 | 7 |
| 6 | 5 | 6 |
| 7 | 4 | 8 |
| 8 | 3 | 10 |
| 9 | 2 | 1 |
| 10 | 1 | 9 |

The dataset is made by only 10 observations and three variables: $x, y, z$.

If we now enter the following command:

```
.  zonotope x y z if x > 3, verbose
```

it will build the zonotope on input variables $x$ and $y$, using variable $z$ as the output variable.
Please notice how only the observations satisfying the condition ($x > 3$) (the first input variable must be greater than 3) will be considered.

Once entered, the last command will display the following result:

---

ZONOTOPE LIBRARY VER 1.3

---

INPUT: SET OF GENERATORS

N. of dimensions (including the output): 3
N. of generators: 7

---

Computation started (it can take a while) ...

---

OUTPUT VECTORS
Diagonal (a row vector):
49 28 46
Tangent between each input generator and the input space:
0.620174
0.896258
0.768221
0.992278
1.17041
0.108465
0.895533

---

OUTPUT STATISTICS
S1: Total volume: 4818
S2: Diagonal norm: 72.808
S3: Sum of squared norms: 867
S4: Gini index: 0.0763405
S5: Tangent of angle btw. diagonal and the input plane: 0.815085
S6: Cosine against output: 0.631799
S7: Cosine of proj. of diagonal on input plane with x axis: 0.868243
S8: Volume against the cube of the norm of the diagonal: 0.0124833

---

Elapsed time (MIN):
0.000000

---

DONE!

---

Now you can display (or reuse) the computed results in this way:

display r(S1) (for the volume)
display r(S2) (for the norm of the diagonal)
...
display r(S8) (volume against the cube of the norm of the diagonal)

matrix list diagonal (for the vector containing the diagonal)
matrix list tangents (for the vector with the tangents)
matrix list gen (for the actually used generators, after if/in)

---

As it will be explained at greater detail in Section 5, building the zonotope of a set of generators has an exponential complexity (i.e., it is very time consuming, especially in high dimension and when the number of generators in high). Therefore, we decided to implement it in C++, in order to lowering as much as possible the running time. In Appendix B we provide a step-by-step instruction on how to compile the C++ source code in order to create the plugin (the latter is a binary file, with extension .plugin). More precisely, we have created an ADO command called `zonotope.ado` which, loads and then calls the C++ Stata plugin.

## Acknowledgments

We gratefully acknowledge Gianluigi Tiesi (Netfarm s.r.l.) for support in creating the git repository and the help in generating the Stata plugin. This project has received financial support from the European Union's Horizon 2020 research and innovation program under grant agreement No. 649186 - ISIGrowth. This paper was presented at the 2019 Italian Stata Users Group meeting (Florence). We thank the participants at the conference for their comments. Suggestions and comments by an anonymous referee were much appreciated. The usual caveat applies. Le Li is the corresponding author.

## D References

Andreoli, F., and C. Zoli. 2014. Measuring Dissimilarity. Technical report, University of Verona, Department of Economics.

Aruka, Y. 2017. Some new perspectives on the inter-country analysis of the world production system. *Evolutionary and Institutional Economics Review* 14(2): 467–498.

Badunenko, O., and P. Mozharovskyi. 2016. Nonparametric frontier analysis using Stata. *Stata Journal* 16(3): 550–589(40).

Baily, M. N., C. Hulten, and D. Campbell. 1992. Productivity Dynamics in Manufacturing Establishments. *Brookings Papers on Economic Activity: Microeconomics* 4: 187–249.

Baldwin, J. R., and M. Rafiquzzaman. 1995. Selection versus evolutionary adaptation: Learning and post-entry performance. *International Journal of Industrial Organization* 13(4): 501–522.

Bartelsman, E. J., and M. Doms. 2000. Understanding Productivity: Lessons from Longitudinal Microdata. *Journal of Economic Literature* 38(2): 569–594.

Belotti, F., S. Daidone, G. Ilardi, and V. Atella. 2013. Stochastic frontier analysis using Stata. *Stata Journal* 13(4): 719–758(40).

Cazals, C., J.-P. Florens, and L. Simar. 2002. Nonparametric frontier estimation: a robust approach. *Journal of econometrics* 106(1): 1–25.

Charnes, A., W. W. Cooper, and E. Rhodes. 1978. Measuring the Efficiency of Decision Making Units. *European Journal of Operations Research* 2: 429–444.

Daraio, C., and L. Simar. 2005. Introducing environmental variables in nonparametric frontier models: a probabilistic approach. *Journal of productivity analysis* 24(1): 93–121.

———. 2007. Conditional nonparametric frontier models for convex and nonconvex technologies: a unifying approach. *Journal of Productivity Analysis* 28(1): 13–32.

Disney, R., J. Haskel, and Y. Heden. 2003. Entry, Exit and Establishment Survival in UK Manufacturing. *Journal of Industrial Economics* 51(1): 91–112.

Dosi, G. 2007. Statistical Regularities in the Evolution of Industries. A Guide through some Evidence and Challenges for the Theory. In *Perspectives on Innovation*, ed. F. Malerba and S. Brusoni. Cambridge University Press.

Dosi, G., and M. Grazzi. 2006. Technologies as problem-solving procedures and technologies as input–output relations: some perspectives on the theory of production. *Industrial and Corporate Change* 15(1): 173–202.

Dosi, G., M. Grazzi, L. Li, L. Marengo, and S. Settepanella. forthcoming. Productivity Decomposition in Heterogeneous Industries. *The Journal of Industrial Economics* .

Dosi, G., M. Grazzi, L. Marengo, and S. Settepanella. 2016. Production Theory: Accounting for Firm Heterogeneity and Technical Change. *The Journal of Industrial Economics* LXIV: 875–907.

Eurostat. 2008. NACE Rev. 2 Statistical classification of economic activites in the European Community. Methodologies and working papers, Eurostat.

Farrell, M. J. 1957. The Measurement of Productive Efficiency. *Journal of the Royal Statistical Society. Series A (General)* 120(3): 253–290.

Franciosi, M., S. Settepanella, and A. Terni. 2020. The robustness of the generalized Gini index. arXiv:2007.12924.

Griliches, Z., and J. Mairesse. 1999. Production Functions: The Search for Identification. In *Econometrics and Economic Theory in the Twentieth Century: the Ragner Frisch Centennial Symposium*, ed. S. Steiner. Cambridge University Press: Cambridge.

Hildenbrand, W. 1981. Short-Run Production Functions Based on Microdata. *Econometrica* 49(5): 1095–1125.

Ji, Y., and C. Lee. 2010. Data envelopment analysis. *Stata Journal* 10(2): 267–280(14).

Koopmans, T. C. 1977. Examples of Production Relations Based on Microdata. In *The Microeconomic foundations of macroeconomics*, ed. G. C. Harcour. London: Macmillan Press.

Koshevoy, G., and K. Mosler. 1996. The Lorenz Zonoid of a Multivariate Distribution. *Journal of the American Statistical Association* 91(434): 873–882.

———. 1997. Multivariate gini indices. *Journal of Multivariate Analysis* 60(2): 252–276.

Mosler, K. 1994. Majorization in economic disparity measures. *Linear algebra and its applications* 199: 91–114.

Simar, L., and V. Zelenyuk. 2011. Stochastic FDH/DEA estimators for frontier analysis. *Journal of Productivity Analysis* 36(1): 1–20.

Syverson, C. 2011. What Determines Productivity? *Journal of Economic Literature* 49(2): 326–65.

The Namibia Statistics Agency. 2013. Namibia household Income and Expenditure Survey 2009/2010 (NAM_2009_HIES_v01_M_v01_A_PUF). Data retrieved from Namibia Statistics Agency, `https://nsa.org.na/microdata1/index.php/catalog/6/study-description` on June 20, 2019.

United Nations. 2008. International Standard Industrial Classification of All Economic Activities Revision 4. Statistical papers series m no. 4/rev.4, UN - Department of Economic and Social Affairs.

#### About the authors

Marco Cococcioni is Associate Professor at the Department of Information Engineering at University of Pisa, Italy. He obtained the PhD from the same university in 2004.

Marco Grazzi is Professor at Department of Economic Policy at Università Cattolica del Sacro Cuore, Milano, Italy. He earned his PhD at Scuola Superiore Sant'Anna, Italy, in 2006.

Le Li is Associate Professor at School of Economics at Guangzhou College of Commerce, China. He got his PhD in Economics from Scuola Superiore Sant'Anna, Italy, in 2016.

Federico Ponchio is researcher at ISTI within the Italian CNR where he is affiliated at the Visual Computing Laboratory. After graduating in Math at the Scuola Normale Superiore, Pisa. He got his PhD at Clausthal TU.