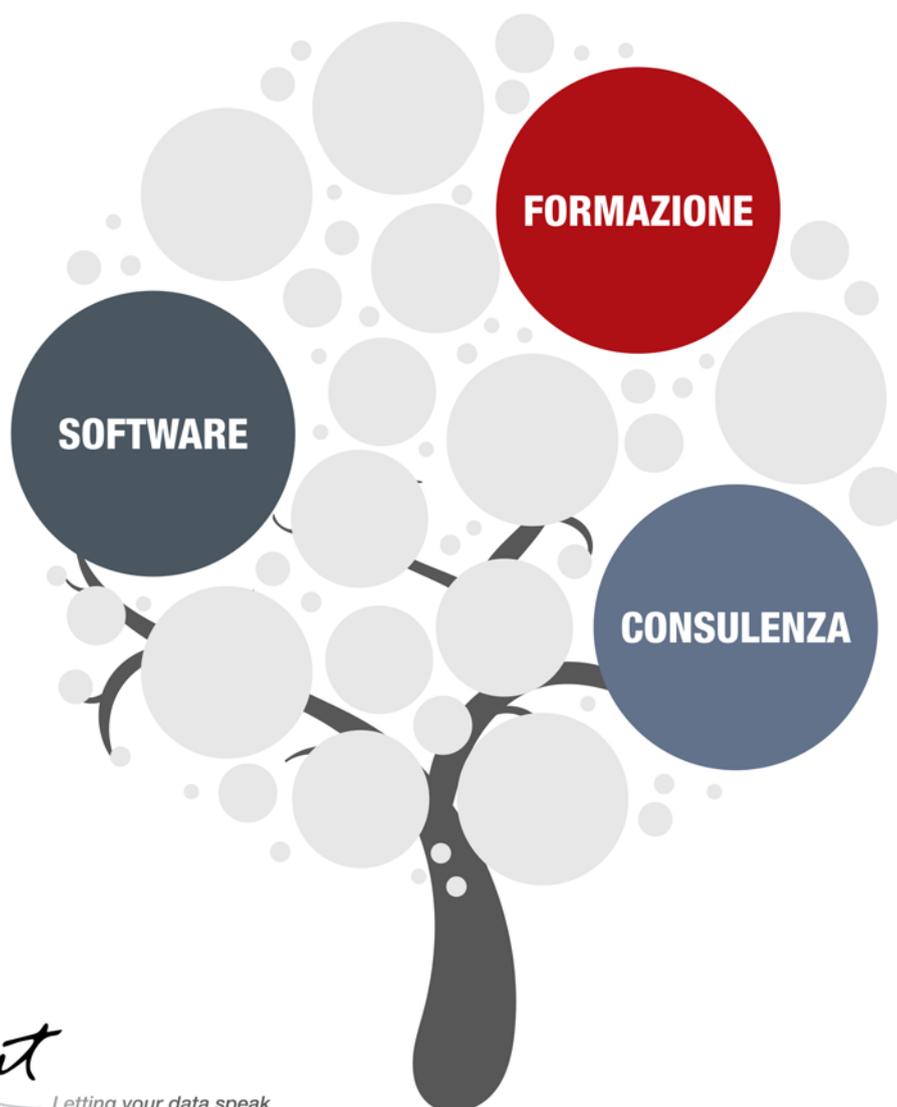


Measuring Heterogeneity and Efficiency of Firms within the same Industry: a C++ plugin for Stata for computing the Zonotope

*Marco Cococcioni** Pisa University, Marco Grazzi Catholic University of Sacro Cuore, Le Li Chuo University Tokyo, Federico Ponchio CNR Pisa

XVI Italian Stata Users Group Meeting
Firenze, 26-27 September 2019

14.00 - 14.50 SESSION IV - COMMUNITY CONTRIBUTED, II





Measuring Heterogeneity And Efficiency of Firms within the Same Industry: a C++ plugin for Stata for computing the *Zonotope*

Marco Cococcioni, Marco Grazzi, Le Li, Federico Ponchio

26 September 2019

XVI Italian Stata Users Group Meeting – Florence (Italy)

email: marco.cococcioni@unipi.it

Introduction

In this work we describe the new Stata command **zonotope**, which:

- provides a measure of productivity that fully accounts for the existing heterogeneity across firms within the same industry;

- allows to assess the extent of multi-dimensional heterogeneity with applications to production analysis and productivity measurement;

- is based on pure geometric concepts.

After describing how to compute the **zonotope** geometrically, we will show how to use the **zonotope** command to perform new empirical analyses.

Empirical Analysis in Economics: the traditional approach

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

The ubiquitous presence of heterogeneity

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.’

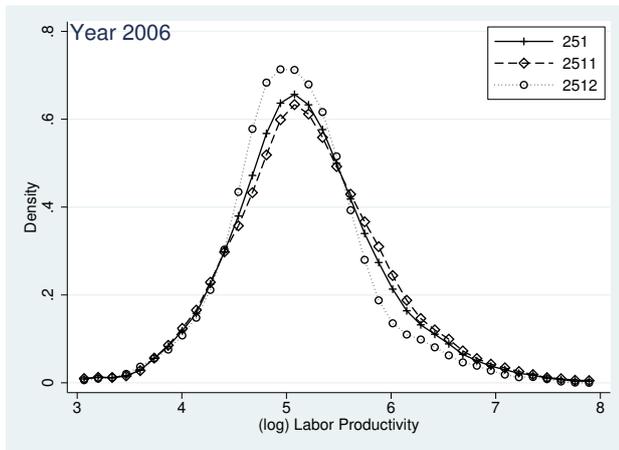
But something like Mandelbrot’s fractal phenomenon seems to be at work here also: the observed variability-heterogeneity does not really decline as we cut our data finer and finer.

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.

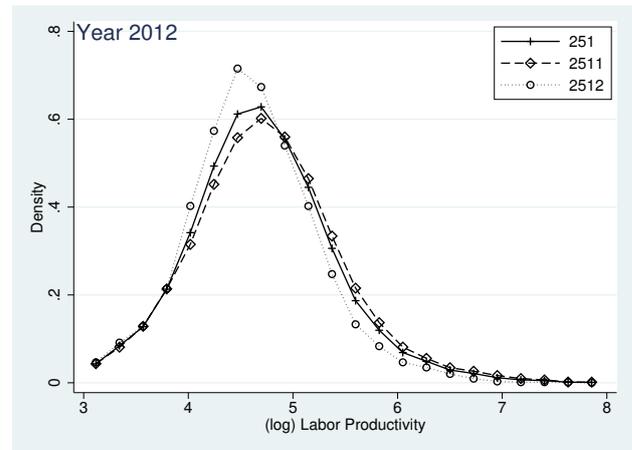
The Persistent Heterogeneity Phenomenon 1/3

To illustrate the phenomenon of persistent heterogeneity, in Figure 1 we provide some empirical evidence focusing on the labor productivity distribution.

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 the next.



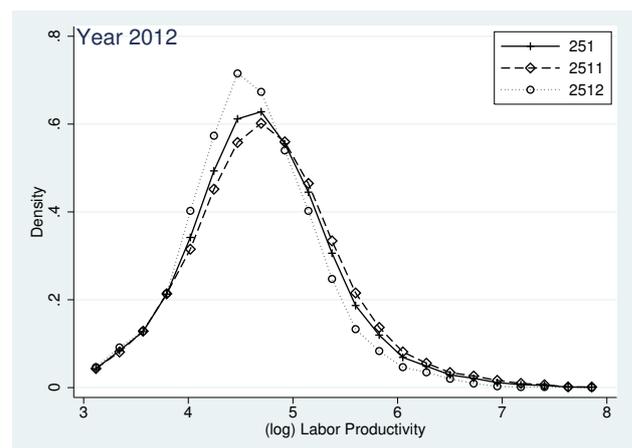
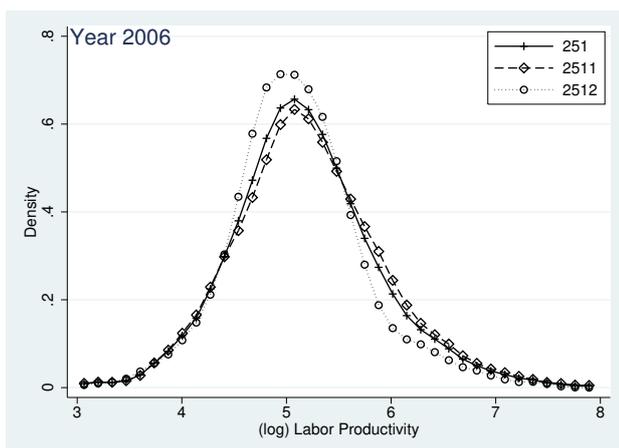
(a) Year 2006



(b) Year 2012

Figure 1: Empirical Distribution of (log) labor productivity in Sector NACE 251 in Italy, and two nested sectors NACE 2511 and NACE 2512.

The Persistent Heterogeneity Phenomenon 2/3

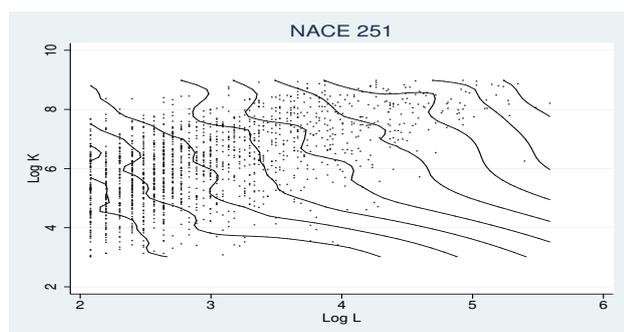


In the left figure, we can see that the productivity distribution at 3-digit level in 2006, i.e. the solid line, is sufficiently widespread to indicate the huge productivity gap between the most productive firms and the least productive ones.

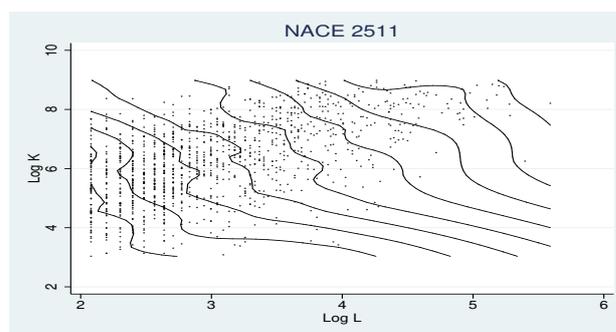
Further this heterogeneity does not disappear when we focus on more similar firms, moving from the 3-digit, to the 4-digit industrial classification, i.e. the dashed and dotted line; we still observe significantly different productivity levels among firms.

The persistence of heterogeneity indicates not only that heterogeneity holds when increasing the level of dis-aggregation but also that it holds over time.

The Persistent Heterogeneity Phenomenon 3/3



(a) NACE 251, Italy, 2006



(b) NACE 2511, Italy, 2006

The left figure provides a representation of production activities of firms within a given 3-digit industry 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 (proxied by turnover). 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.

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 (NACE 2511) as reported in the right figure above.

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.

Production Activity Representation 1/2

Similar to Koopmans (1977); Hildenbrand (1981), and Dosi et al. (2016), we denote the production activity, as representing the actual technique of production unit i , by a vector

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

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 $(\alpha_{i_1}, \dots, \alpha_{i_l})$ 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

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

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 $\{a_i\}_{1 \leq i \leq N}$.

More explicitly, the *short-run feasible industry production function* as the Zonotope

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

Production Activity Representation 2/2

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

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

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

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

This definition implies that given the level u_1, \dots, u_l of inputs for the industry, the maximum total output could be achieved by allocating, without any restrictions, the amounts u_1, \dots, 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).

The Zonotope Diagonal d_Y

Within the Zonotope framework, Dosi et al. (2016) defined the main diagonal of a Zonotope Y as the line joining the origin $O = (0, \dots, 0)$ with its opposite vertex in Y .

They called this diagonal the **production activity of the industry**, since it expresses both the amount of inputs employed and output produced by the industry.

Its definition is very simple, since it is the sum of individual production activities of the N production units involved in the industry, i.e.

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

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 a zero volume (**minimum heterogeneity case**).

On the other hand, the **maximum 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**.

From this simple observation on these two extreme cases, it is possible to derive a rigorous measure for industry heterogeneity (see next slide).

A new measure of **heterogeneity** definable on Zonotopes

Let $A_{i_1, \dots, i_{l+1}}$ be the matrix whose rows are vectors $\{\alpha_{i_1}, \dots, \alpha_{i_{l+1}}\}$ and $\Delta_{i_1, \dots, i_{l+1}}$ be its determinant.

It is well known that the volume of the zonotope Y in \mathbb{R}^{l+1} is given by:

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

where $|\Delta_{i_1, \dots, i_{l+1}}|$ is the module of the determinant $\Delta_{i_1, \dots, i_{l+1}}$.

However, the value of $Vol(Y)$ depends on both **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 production activities $\{a_i\}_{1 \leq i \leq N}$ is divided by the volume of the **parallelotope** having diagonal d_Y .

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 $Vol(P_Y)$ is its volume).

The normalized volume, $G(Y)$, is named the **Gini volume**.

A new measure of **productivity** definable on Zonotopes

The angle formed by the industry production activity vector d_Y with the space generated by all inputs expresses the industry productivity.

The tangent of this angle can be an appropriate measure for it.

More precisely, the measure of productivity P for a given industry including N firms at current period, is

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

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

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

$\Theta_j(v)$ represents the angle formed by the vector v and the space generated by all entities in vector $pr_{-j}(v)$

$\|v\|$ represents the normal of the vector v .

The Zonotope Stata command

In this section we introduce the Stata command **zonotope**. Also note that the software is available at <https://github.com/zonotopes> and can be further developed or customized.¹

Syntax The syntax of the command to compute the zonotope is:

```
zonotope varlist [if exp in range] [, verbose]
```

where the option `verbose`, when used, causes the program to print on screen all the computed quantities (see below).

varlist is a list of vector-valued variables, all of the same length N . The i -th value on each vector is associated to the i -th generator a_i .

The first l variables are the input of the relation we want to model, while the last $(l + 1)$ is the output variable.

¹The software that we are presenting here is also available for C++, Matlab and R.

Outputs and return values 1/5

The zonotope command returns two vectors: *diagonal* and *tangents*. In addition, it returns a number of scalar values, such as

`nrow`: the number of generators actually used;

`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 $S1, S2, \dots, S8$ 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.

The variables returned by zonotope can be accessed using the `r()` command. As an example, the elapsed time can be displayed using the command

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

Outputs and return values 2/5

Output vector: Diagonal

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

To be specific,

$$d_Y = (\beta_1, \dots, \beta_l, \beta_{l+1}) = \left(\sum_{i=1}^N \alpha_{i1}, \dots, \sum_{i=1}^N \alpha_{il}, \sum_{i=1}^N \alpha_{i,l+1} \right).$$

Clearly, it is an $(l + 1)$ -dimensional (row) vector.

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.

The eight statistics mentioned above are:

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).

Outputs and return values 3/5

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.

$$S2 \equiv \|d_Y\| = \sqrt{\sum_{i=1}^{I+1} \beta_i^2}.$$

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,

$$S3 = \sum_{i=1}^N \left(\sum_{j=1}^{I+1} \alpha_{ij}^2 \right).$$

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,

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

where P_Y denote the parallelotope with diagonal d_Y .

Outputs and return values 4/5

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,

$$S5 \equiv \text{tg}(\Theta_{I+1}(d_Y)) = \frac{\beta_{I+1}}{\sqrt{\sum_{j=1}^I \beta_j^2}}.$$

Statistic S6: Cosine against output

$S6$ reports the cosine of the complementary angle of $\Theta_{I+1}(d_Y)$ as

$$S6 = \frac{\beta_{I+1}}{\sqrt{\sum_{j=1}^{I+1} \beta_j^2}}.$$

(due to the complementary angle relationship, the following relation between $S5$ and $S6$ also holds: $S5 = \frac{S6}{\sqrt{1-(S6)^2}}$)

Outputs and return values 5/5

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

$$S7 = \frac{\beta_1}{\sqrt{\sum_{j=1}^I \beta_j^2}} .$$

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,

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

An example of usage

As an example, by entering the following commands at the Stata prompt:

```
. use http://www.stata-press.com/data/r11/test
. 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 what is shown in next slide.

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 enter:

display r(S1) (for the volume)

display r(S2) (for the norm of the diagonal), etc.

3D Visualization

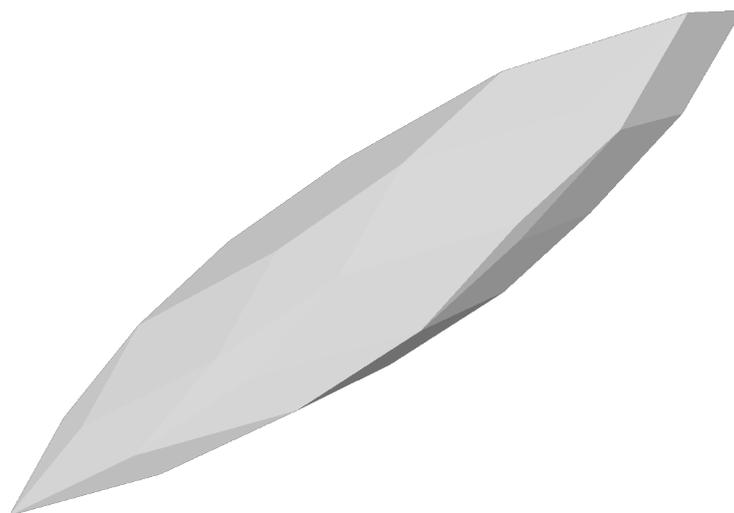


Figure 2: 3D shape of the 3D zonotope (3 variables: 2 inputs and 1 output) with 7 generators on which the **zonotope** command has been called. Picture created using MeshLab.

zonotope as a C++ plugin within an .ado file

As it will be shown in the subsequent slides, 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 is high).

Therefore, we decided to implement it in C++, in order to lower as much as possible the running time.

In Appendix C 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.

Test on real-world datasets 1/5

In this section, using the **zonotope** command, we provide two different examples related to performing economic investigations.

First, based on firm-level data, we compute the heterogeneity and productivity levels of industry as suggested by Dosi et al. (2016).

Second, firm-level measures of productivity are computed resorting to **zonotope** command in Stata.

Used Data

To perform the empirical analysis, we employ firms' data from **Amadeus** dataset provided by Bureau van Dijk (hereafter BvD).

The BvD physical media (DVD Blu-Ray, October 2015) contains comprehensive information of balance sheets and income statements on around 21 million companies across Europe and covers the period between 2004 and 2013.

Amadeus also provides, for each firm, its 4-digit NACE² codes at 2012, which allows us to classify all firms into different industrial sectors at different disaggregate level (to the maximum 4-digit).

²NACE is The Statistical Classification of Economic Activities in the European Community

Test on real-world datasets 2/5

Table 1 reports the industries we selected.

For each firm in the selected industries, number of employees and fixed assets are chosen to proxy for its inputs and turnover value for its output.

All of these values, except for the number of employees, are measured in thousands euro and deflated at 4-digit NACE level with year 2010 as benchmark year to perform inter-temporal comparison.

Table 1: List of Selected Industries

NACE	Name of the Industry
1011	Processing and preserving of meat
1091	Manufacture of prepared feeds for farm animals
1105	Manufacture of beer
1310	Preparation and spinning of textile fibres
1411	Manufacture of leather clothes
1920	Manufacture of refined petroleum products
2011	Manufacture of industrial gases
2012	Manufacture of dyes and pigments
2013	Manufacture of other inorganic basic chemicals
2014	Manufacture of other organic basic chemicals
2015	Manufacture of fertilisers and nitrogen compounds
2016	Manufacture of plastics in primary forms
2211	Manufacture of rubber tyres and tubes; retreading and rebuilding of rubber tyres
2221	Manufacture of plastic plates, sheets, tubes and profiles
2351	Manufacture of cement
2352	Manufacture of lime and plaster
2451	Casting of iron

Test on real-world datasets 3/5

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** (i.e., the Gini index) and the **tangent value of the angle formed by industry production vector and its input plane** are easily computed as **S4** and **S5** reported in our package.

We can use these two numbers to measure the industry heterogeneity and productivity level for one specific industry at a given point in time.

Columns (2) and (3) of the table in next slide (Table 2) report these two results for selected industries in 2006, while column (1) reports the number of firms within that industry-year cell.

Test on real-world datasets 4/5

Table 2: Computation Results for Gini Coefficient and Productivity among Selected Sectors and Years in Italy - R^3 Case

NACE	Year 2006			Year 2009			Year 2012		
	Obs (1)	Gini (2)	tg (3)	Obs (4)	Gini (5)	tg (6)	Obs (7)	Gini (8)	tg (9)
1011	160	0.202	6.211	222	0.234	4.235	346	0.261	4.492
1091	115	0.124	6.514	134	0.115	6.876	162	0.138	9.240
1105	14	0.014	1.552	31	0.020	1.783	70	0.056	1.420
1310	386	0.152	2.978	431	0.190	2.064	603	0.199	2.922
1411	57	0.092	5.593	69	0.083	3.099	142	0.113	4.535
1920	142	0.153	4.066	168	0.164	2.683	212	0.259	5.807
2011	43	0.098	0.877	52	0.111	0.923	68	0.128	1.264
2012	47	0.092	3.425	53	0.098	2.294	69	0.119	3.066
2013	38	0.074	1.477	41	0.084	1.047	47	0.080	2.260
2014	18	0.083	1.285	26	0.091	1.070	40	0.208	2.046
2015	65	0.105	3.802	85	0.168	3.248	126	0.173	3.470
2016	177	0.207	3.092	189	0.152	2.151	259	0.163	3.150
2211	54	0.039	2.445	58	0.060	1.495	82	0.077	1.878
2221	225	0.138	3.362	245	0.125	2.228	368	0.130	2.753
2351	28	0.002	0.924	33	0.002	0.757	45	0.001	0.751
2352	45	0.027	1.343	44	0.027	0.879	62	0.043	0.883
2451	86	0.070	3.682	89	0.088	1.473	121	0.092	2.307

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 NACE 1310 while 160 firms in NACE 1011.

Test on real-world datasets 5/5

However given a larger number of firms, we don't expect that Gini coefficient, i.e. the normalized volume of zonotope, of NACE 1310 to be necessarily bigger than that of NACE 1011. Indeed, based on our data sample, the Gini coefficient of NACE 1011 is larger, 0.202 as compared to 0.152 of NACE 1310.

After taking into account the effect from the number of firms, we notice that among different industries, the heterogeneity levels are different. For example, in 2006 the Gini coefficients vary from 0.002 for NACE 2351 to 0.207 for NACE 2016. 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 NACE 1011 increases from 0.202 in 2006 to 0.234 in 2009, and to 0.261 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.

Example: computing Firm Productivity 1/2

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 , is,

$$p_i = \text{tg}(\Theta_{l+1}(a_i)) = \frac{\alpha_{l+1}}{\|pr_{-(l+1)}(a_i)\|} \quad (6)$$

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

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 Italian firms from industry NACE 2014 in 2006 and report them in the column "Year 2006" of table 3.

Example: computing Firm Productivity 2/2

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.

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.

Table 3: Productivity of Italian Firms from Industry NACE 2014 - R^3 Case

Firm ID	Year 2006	Year 2009	Year 2012
1	1.436	2.504	4.540
2	6.841	6.924	15.153
3	0.373	0.186	0.307
4	5.951	4.560	7.639
5	0.854	0.609	0.982
6	5.559	3.140	5.884
7	11.764	14.534	15.495
8	22.626	10.967	4.343
9	14.394	18.090	3.467
10	4.706	5.559	11.455
11	0.479	0.594	0.745
12	1.036	1.059	1.511

Analysis of Zonotope Computing Time 1/2

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 $(l + 1)$.

Next table (Table 4) 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, the subsequent one (Table 5) 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.

Analysis of Zonotope Computing Time 2/2

Table 4: 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 5: 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

Conclusions 1/2

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 of empirical observations.

In this respect the **zonotope** package provides a rigorous way to perform empirical analysis taking advantage of 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.

Conclusions 2/2

Such differences do not vanish over time: selection, if at works, 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. Several versions of the code are provided, Stata, R, Matlab, C++ and they have been organized so as to facilitate further development, also from neighboring fields (computational geometry, computer graphics, machine learning, etc.).

Appendix A - The zonotope help

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

Syntax
`zonotope varlist [if expr] [in range] [, verbose]`

Description
This function builds the zonotope from a list of (l+1) vector variables, where the first l are the input ones and the (l+1)-th the output one.
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, and two vectors: r(diagonal) and r(tangents).
Finally, the zonotope command prints on screen all the computed quantities, when it is called with the 'verbose' option.

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 zonotope_demo`
- `. run zonotope_demo2`

Appendix 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
```

Then run:

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

To run the demo, execute the following do file:

```
. run zonotope_demo
```

If the demo runs correctly, you are done.

Otherwise, the plugin must be generated from scratch for your operating system, as explained in the next two slides.

How to generate the zonotope plugin 1/2

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.

How to generate the zonotope plugin 2/2

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
- on MacOS 10.12.3 (16D32), 64 bit, using LLVM C++ compiler.

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

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.

We also gratefully acknowledge for Namibia Statistics Agency for providing data from National Household Income and Expenditure Survey (2009-2010).

This project has received financial support from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 649186 - ISIGrowth.

M. Grazzi gratefully acknowledges financial support from the University of Bologna, internal grant RFO14GRAZM.

References

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