Workshop
ADVANCES IN DATA SCIENCE FOR BIG AND COMPLEX DATA
From data to classes and classes as new statistical units

UNIVERSITY PARIS-DAUPHINE
January 10-11, 2019
ROOM C108 First floor

Symbolic input output analysis: harmonic analysis approach to combine statistical distributions

Andrej Srakar, PhD, PhDc

Institute for Economic Research (IER) and Faculty of Economics, University of Ljubljana

Research problem

- Input-output (IO) analysis is a non-stochastic approach, prone to numerous limitations – it is usually described as a rather „crude“ tool
- In the literature there have been several approaches to including the stochastic element in IO analysis (e.g. West, 1986; Jansen, 1994; Ten Raa, 2005; Sancho et al., 2011; 2012; 2015; 2017; Lenzen et al., 2014)
- We provide a completely new one, which has potential of significantly changing the field and calculations, providing them a) stochastic element, so they can be easier complemented in e.g. regression analysis; b) better accuracy and predictability purposes by including significantly more information on the cells in IO tables
- We test the approach on some preliminary/pilot datasets

Structure of the presentation

- Input output analysis and its flaws
- SDA and the idea of the paper
- Calculus of distributions
- Derivation of the new Leontief formulas and multipliers
- Some empirical results
- Generalization to CoDA and FDA
- Conclusion

Input-output analysis basics

Input ↓ Output→	Agriculture	Industry	Services	Total outputs
Agriculture	2,180	81,687	1,143	200,345
Industry	27,709	98,036	25,457	538,119
Services	11,020	32,242	19,487	301,311
Total inputs	200,345	538,119	301,311	1,945,233

Table 1: Fragment of an input-output table

Input ↓ Output→	1	2	3	Total final demand	Total outputs
1	x_{11}	x_{12}	x_{13}	Y_1	X_1
2	x_{21}	x_{22}	x_{23}	Y_2	X_2
3	x_{31}	x_{32}	x_{33}	Y_3	X_3
All primary inputs	Z_1	Z_2	Z_3	-	-
Total inputs	X_1	X_2	X_3	-	-

Table 2: Fragment of an input-output table

Input-output analysis basics

$$X_1 = x_{11} + x_{12} + x_{13} + Y_1 \tag{6}$$
$$X_2 = x_{21} + x_{22} + x_{23} + Y_2 \tag{7}$$
$$X_3 = x_{31} + x_{32} + x_{33} + Y_3 \tag{8}$$
$$a_{ij} = \frac{x_{ij}}{X_j} \tag{9}$$
$$i \ldots \text{row} \tag{10}$$
$$j \ldots \text{column} \tag{11}$$

$$x_{ij} = a_{ij} X_j \tag{12}$$
$$\Rightarrow X_1 = a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + Y_1 \tag{13}$$
$$\Rightarrow X_2 = a_{21}X1 + a_{22}X_2 + a_{23}X_3 + Y_2 \tag{14}$$
$$\Rightarrow X_3 = a_{31}X_1 + a_{32}X_2 + a_{33}X_3 + Y_3 \tag{15}$$

$$(I - A)X = Y \tag{17}$$
$$X = (I - A)^{-1}Y \tag{18}$$

Some input-output analysis limitations

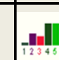
- Its framework rests on Leontief's basic assumption of constancy of input co-efficient of production.
- Ignores the possibility of factor substitution.
- The assumption of linear equations, which relates outputs of one industry to inputs of others.
- The rigidity of the input-output model cannot reflect such phenomena as bottlenecks, increasing costs, etc.
- The purchases by the government and consumers are taken as given and treated as a specific bill of goods.
- There is no mechanism for price adjustments.
- The use of capital in production necessarily leads to stream of output at different points of time being jointly produced.

Symbolic Data Analysis

Standard data table describing individuals.

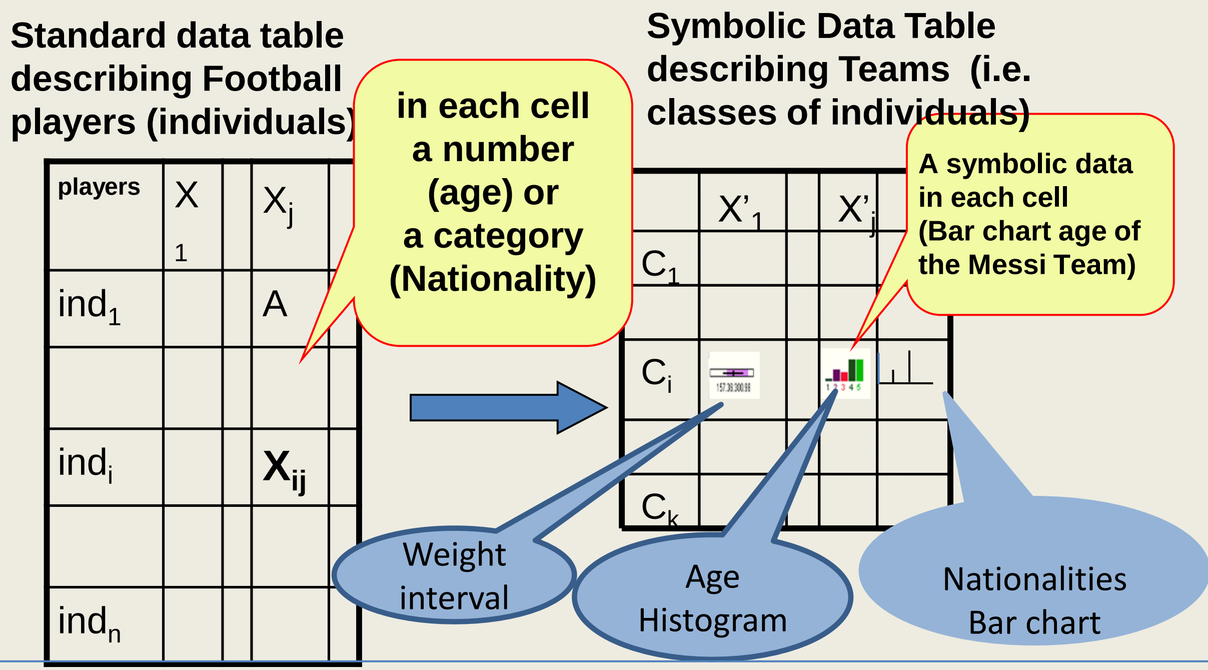
players	X_1		X_j	
ind_1				
ind_i			$\mathbf{X_{ij}}$	
ind_n				

Symbolic Data Table describing classes obtained from a symbolic representation process

	X'_1		X'_j	
C_1				
C_i				
C_k				

Symbolic Data Analysis

From standard random variables to random variables of random variable value



In that way, we obtain new kinds of data expressing variability inside classes and called "symbolic" as they cannot be reduced to numbers without losing much information.

Symbolic input-output analysis: the idea

- Instead of a fixed, gross, aggregate number in the cell of intermediary production, we include a *distribution* (quantiles, constructed from the data of legal subjects)
- The formulas for constructing the (production) multipliers significantly change as we are now in the „world“ of combining distributions – algebra of random variables
- We gain a stochastic element and significantly more information on the distribution of the intermediary production between sectors
- Additional: *IO analysis can also be done for compositional and functional data cells in IO tables*

Calculus of distributions – four operations

- **Addition: *convolution, C*** – a mathematical operation on two functions to produce a third function, giving the integral of the pointwise multiplication of the two functions as a function of the amount that one of the original functions is translated

$$(f * g)(z) = \int_{-\infty}^{\infty} f(x)g(z-x)dx = \int_{-\infty}^{\infty} f(z-x)g(x)dx = (g * f)(z)$$

- **Generalized convolution** (Seong Kang et al., 2010)**:**

Let $\xi_1, \xi_2, \ldots, \xi_n$ be independent and identically distributed random variables with the common distribution function F and probability density function f. Then the distribution function of the sum ζ_n is the n-fold convolution of itself F such as

$$F^{n*}(x) = F^{(n-1)*}(x) * F(x) \quad (n \geq 2)$$

where $F^{1*}(x) = F(x)$ and its probability density function is

$$f^{n*}(x) = f^{(n-1)*}(x) * f(x) \quad (n \geq 2)$$

where $f^{1*}(x) = f(x)$.

- ***Difference, D***:

$$f(z = y - x) = -\int_{-\infty}^{\infty} g(x)h(y)dy$$

Some common knowledge convolutional relations

- $\sum_{i=1}^{n} Bernoulli(p) \sim Binomial(n, p) \quad 0 < p < 1 \qquad n = 1,2,\dots$
- $\sum_{i=1}^{n} Binomial(n_i, p) \sim Binomial(\sum_{i=1}^{n} n_i, p) \quad 0 < p < 1 \quad n_i = 1,2,\dots$
- $\sum_{i=1}^{n} NegativeBinomial(n_i, p) \sim NegativeBinomial(\sum_{i=1}^{n} n_i, p) \;\; 0 < p < 1 \qquad n_i = 1,2,\dots$
- $\sum_{i=1}^{n} Geometric(p) \sim NegativeBinomial(n, p) \;\; 0 < p < 1 \;\; n = 1,2,\dots$
- $\sum_{i=1}^{n} Poisson(\lambda_i) \sim Poisson(\sum_{i=1}^{n} \lambda_i) \qquad \lambda_i > 0$
- $\sum_{i=1}^{n} Normal(\mu_i, \sigma_i^2) \sim Normal\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right) \;\; -\infty < \mu_i < \infty \quad \sigma_i^2 > 0$
- $\sum_{i=1}^{n} Cauchy(a_i, \gamma_i) \sim Cauchy(\sum_{i=1}^{n} a_i, \sum_{i=1}^{n} \gamma_i) \;\; -\infty < a_i < \infty \quad \gamma_i > 0$
- $\sum_{i=1}^{n} Gamma(\alpha_i, \beta) \sim Gamma(\sum_{i=1}^{n} \alpha_i, \beta) \qquad \alpha_i > 0 \quad \beta > 0$
- $\sum_{i=1}^{n} Exponential(\theta) \sim Gamma(n, \theta) \qquad \theta > 0 \quad n = 1,2,\dots$
- $\sum_{i=1}^{n} \chi^2(r_i) \sim \chi^2(\sum_{i=1}^{n} r_i) \quad r_i = 1,2,\dots$
- $\sum_{i=1}^{r} N^2(0,1) \sim \chi_r^2 \qquad r = 1,2,\dots$
- $\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \sigma^2 \chi_{n-1}^2 \; where \;\; X_1, \dots, X_n \; is\; a\; random\; sample\; from\; N(\mu, \sigma^2)$

Calculus of distributions – four operations

- **Multiplication: *product distribution, P***

- While the probability density function (PDF) of the sum of two independent random variables is easily described as the convolution of their PDFs, the expression for the PDF of the product is more complicated.
- Rohatgi (1976): Let X and Y be continuous random variables with joint PDF $f_{X,Y}(x,y)$. The PDF of $V = XY$ is

$$f_{XY}(v) = \int_{-\infty}^{\infty} f_{X,Y}\left(x, \frac{v}{x}\right) \frac{1}{|x|} dx$$

For the easiest case of Gaussian variables, it is possible to use the following theorem (Beylkin, Monzón and Satkauzkas, 2016):

- The PDF of the product of two independent random variables X and the Gaussian variable $Y \sim N(\mu_y, \sigma_y^2)$ can be approximated using approximate multiresolution analysis as

$$\left| p(t) - \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} w_j^k \phi_{jk}(t) \right| \leq \epsilon p(t)$$

where

$$w_j^k = 2^{-j/2} log2 \int_0^1 \frac{w_+(\tau) + w_-(\tau)}{\sqrt{1-4^{\tau-2}}} d\tau$$

$$w_+(\tau) = \frac{1}{\sqrt{2\pi}\sigma_y} f(\frac{2^{2-j}}{\sqrt{2\alpha}\sigma_y} 2^{-\tau}) e^{-\frac{\alpha 4^{\tau-2}}{1-4^{\tau-2}}(k - \frac{\mu_y}{\sqrt{2\alpha}\sigma_y} 2^{-\tau+2})^2}$$

$$w_-(\tau) = \frac{1}{\sqrt{2\pi}\sigma_y} f(-\frac{2^{2-j}}{\sqrt{2\alpha}\sigma_y} 2^{-\tau}) e^{-\frac{\alpha 4^{\tau-2}}{1-4^{\tau-2}}(k + \frac{\mu_y}{\sqrt{2\alpha}\sigma_y} 2^{-\tau+2})^2}$$

- Beylkin, Monzón and Satkauzkas also derive similar results for combining Gaussian with Cauchy, Laplace and Gumbel variables

Calculus of distributions – four operations

- **Division: *ratio distribution, R***

$$p_Z(z) = \int_{-\infty}^{\infty} |y| p_{X,Y}(zy, z) dy$$

- Also suggested: Mellin transform.
- The algebraic rules known with ordinary numbers do not apply for the algebra of random variables. For example, if a product is *C = AB* and a ratio is *D=C/A* it does not necessarily mean that the distributions of *D* and *B* are the same.

- **Inverse distribution**:
- In general, given the probability distribution of a random variable *X* with strictly positive support, it is possible to find the distribution of the reciprocal, *Y = 1 / X*. If the distribution of *X* is continuous with density function *f(x)* and cumulative distribution function *F(x)*, then the cumulative distribution function, *G(y)*, and PDF of the reciprocal are found as:

$$G(y) = 1 - F(\frac{1}{y})$$

$$g(y) = \frac{1}{y^2} f(\frac{1}{y})$$

Calculus of distributions – four operations

- A known result (M.D. Springer, 1979; Hazewinkel, 1991; Whittle, 2000):

Random variables have the following properties:

- complex constants are random variables;
- the sum of two random variables is a random variable;
- the product of two random variables is a random variable;
- addition and multiplication of random variables are both commutative; and
- there is a notion of conjugation of random variables, satisfying $(ab)^* = b^*a^*$ and $a^{**} = a$ for all random variables a,b and coinciding with complex conjugation if a is a constant.

*Therefore, the random variables form complex commutative *-algebras (i.e. involutive algebras),* see Wegge-Olsen, 1993; Davidson, 1996; Cuntz and Echterhoff, 2000; Baez, 2015; Weisstein, 2015.

Calculus of distributions – four operations

An *involutive algebra*:

an involutive (*-) ring with involution * that is an associative algebra over a commutative *-ring R with involution ′, such that $(r\,x)^* = r'\,x^*\ \ \forall r \in R, x \in A$

- The base *-ring R is usually the complex numbers (with ′ acting as complex conjugation) and is commutative with A such that A is both left and right algebra.
- Since R is central in A, that is, $rx = xr$, $\forall r \in R, x \in A$, the * on A is conjugate-linear in R, meaning $(\lambda\, x + \mu\, y)^* = \lambda'\, x^* + \mu'\, y^*$ for $\lambda, \mu \in R$, $x, y \in A$.

Derivation of the new Leontief formulas

- Two different cases – total output is fixed or is a distribution itself
- First case – it is fixed

$$X = [I - c(A)]^{-1} Y$$

$$(1 - \mathbf{a}_{ii}) X_i - \sum_{i \neq j} \mathbf{a}_{ij} X_j = Y_i$$

- Second case – it is a distribution

$$X = [I - r(A)]^{-1} Y$$

$$\mathbf{X}_i - \frac{\mathbf{x}_{ii}}{\mathbf{X}_i} \mathbf{X}_i - \sum_{i \neq j} \frac{\mathbf{x}_{ij}}{\mathbf{X}_j} \mathbf{X}_j = \mathbf{Y}_i$$

Input-output analysis basics

$$X_1 = x_{11} + x_{12} + x_{13} + Y_1 \tag{6}$$
$$X_2 = x_{21} + x_{22} + x_{23} + Y_2 \tag{7}$$
$$X_3 = x_{31} + x_{32} + x_{33} + Y_3 \tag{8}$$
$$a_{ij} = \frac{x_{ij}}{X_j} \tag{9}$$
$$i \ldots \text{row} \tag{10}$$
$$j \ldots \text{column} \tag{11}$$

$$x_{ij} = a_{ij} X_j \tag{12}$$
$$\Rightarrow X_1 = a_{11} X_1 + a_{12} X_2 + a_{13} X_3 + Y_1 \tag{13}$$
$$\Rightarrow X_2 = a_{21} X1 + a_{22} X_2 + a_{23} X_3 + Y_2 \tag{14}$$
$$\Rightarrow X_3 = a_{31} X_1 + a_{32} X_2 + a_{33} X_3 + Y_3 \tag{15}$$

$$(I - A)X = Y \tag{17}$$
$$X = (I - A)^{-1} Y \tag{18}$$

Derivation of the new Leontief formulas

- *Final issue – inverting a random matrix* (based on its spectral properties)
- *A* is almost surely invertible whenever its entries are absolutely continuous (Cayley-Hamilton method, Neumann series method, QR method, random matrix methods, LSMR, LSQR, Kaczmarz method – see Trotter, 1984; Silverstein, 1985; Edelman, 1989; Dumitriu and Edelman, 2002), the case of discrete entries is non-trivial.
- For the later case, use the procedure of Rudelson (2008), using ε-net argument for one part of the sphere and conditional probability arguments (the method of Halász, 1975; 1977) for the other.
- Some special cases: inverse of Ginibre ensemble (matrix of i.i.d. random normal variables), inverse of a Wishart and compound Wishart matrix, inverse of Cauchy.
- Some empirical work done with the help of Wolfram Mathematica.

Simulated examples, using Eurostat data

- Basic multiplier calculation (previous own study)

Country	Year	Mult Sci	Mult PubAd	Mult Hea	Mult Edu	Mult SocC	Mult Advert	Mult Creat	Mult Publ
France	2008	1.9323	1.4744	1.3285	1.2981	1.2795	1.9117	1.6940	1.9783
France	2009	2.2229	1.6001	1.4241	1.3658	1.3607	2.0596	1.9245	2.2389
France	2010	2.1380	1.5724	1.3925	1.3424	1.3317	2.0058	1.8581	2.1151
Germany	2008	1.6253	1.4892	1.4369	1.3093	1.4518	1.6346	1.5126	1.8170
Germany	2009	1.5604	1.4597	1.3707	1.3144	1.3703	1.7272	1.4290	1.7955
Germany	2010	1.5693	1.4569	1.3500	1.3099	1.3522	1.7031	1.4189	1.7618
Italy	2008	1.7456	1.5151	1.3913	1.2527	1.5979	2.0493	1.7475	2.1051
Italy	2009	1.7361	1.4841	1.3829	1.2411	1.5797	2.0398	1.7335	2.1269
Italy	2010	1.8231	1.4717	1.3805	1.2411	1.5832	2.0530	1.7114	2.0865
Netherlands	2008	1.5990	1.6075	1.2844	1.3135	1.3306	1.9698	1.8114	1.6926
Netherlands	2009	1.5967	1.6134	1.2687	1.3007	1.3042	1.9644	1.7662	1.6580
Netherlands	2010	1.5848	1.5831	1.2587	1.2944	1.2937	1.9415	1.7444	1.6163
Portugal	2008	1.4708	1.4056	1.5222	1.2137	1.4818	2.1678	1.7880	1.9942
Portugal	2009	1.4765	1.4390	1.5373	1.2449	1.4485	2.1641	1.8257	1.9797
Portugal	2010	1.4236	1.3933	1.5346	1.2420	1.4168	2.1243	1.7670	1.9295

Simulated examples, using Eurostat data

- Symbolic IO – fixed final outputs, Gaussianity, the variables calculated are Wishart

Country	Year	Mult Sci	Mult PubAd	Mult Hea	Mult Edu	Mult Soc C	Mult Advert	Mult Creat	Mult Publish
France	2008	(239.13;20)	(438.63;20)	(180.36;20)	(166.03;20)	(172.62;20)	(819.53;20)	(754.81;20)	(185.68;20)
France	2009	(998.61;20)	(895.73;20)	(601.87;20)	(64.72;20)	(246.45;20)	(127.66;20)	(628.01;20)	(868.95;20)
France	2010	(304.32;20)	(890.18;20)	(92.20;20)	(314.75;20)	(124.32;20)	(797.95;20)	(476.65;20)	(437.19;20)
Germany	2008	(961.42;20)	(270.95;20)	(187.39;20)	(701.65;20)	(903.26;20)	(947.14;20)	(170.87;20)	(906.01;20)
Germany	2009	(624.50;20)	(75.35;20)	(657.43;20)	(498.08;20)	(459.83;20)	(610.81;20)	(973.75;20)	(925.70;20)
Germany	2010	(87.07;20)	(210.24;20)	(934.87;20)	(588.27;20)	(473.92;20)	(46.19;20)	(985.59;20)	(914.43;20)
Italy	2008	(155.54;20)	(29.82;20)	(30.55;20)	(339.57;20)	(268.57;20)	(86.35;20)	(856.23;20)	(457.82;20)
Italy	2009	(617.96;20)	(922.04;20)	(805.62;20)	(213.80;20)	(483.15;20)	(940.93;20)	(626.21;20)	(378.84;20)
Italy	2010	(630.45;20)	(307.44;20)	(693.79;20)	(632.27;20)	(638.20;20)	(380.53;20)	(167.30;20)	(900.43;20)
Netherlands	2008	(676.15;20)	(635.57;20)	(963.38;20)	(129.15;20)	(74.26;20)	(722.59;20)	(921.83;20)	(3.81;20)
Netherlands	2009	(665.32;20)	(599.26;20)	(118.76;20)	(92.20;20)	(305.33;20)	(609.44;20)	(644.45;20)	(369.53;20)
Netherlands	2010	(197.45;20)	(754.65;20)	(42.77;20)	(198.92;20)	(965.20;20)	(89.80;20)	(957.63;20)	(263.62;20)
Portugal	2008	(314.93;20)	(454.95;20)	(184.13;20)	(593.65;20)	(542.40;20)	(526.14;20)	(958.31;20)	(345.61;20)
Portugal	2009	(678.75;20)	(876.53;20)	(441.58;20)	(877.74;20)	(642.29;20)	(182.44;20)	(594.32;20)	(62.21;20)
Portugal	2010	(963.34;20)	(146.26;20)	(349.93;20)	(463.31;20)	(738.89;20)	(255.28;20)	(425.84;20)	(248.17;20)

Simulated examples, using Eurostat data

- Symbolic IO – outputs as distributions, Gaussianity, the variables calculated are Cauchy

Country	Year	Mult Sci	Mult PubAd	Mult Hea	Mult Edu	Mult Soc C	Mult Advert	Mult Creatart	Mult Publish
France	2008	(1.709;0.551)	(1.262;1.482)	(1.850;0.511)	(1.925;1.844)	(1.679;0.779)	(1.225;1.977)	(1.574;1.078)	(1.493;1.529)
France	2009	(1.345;0.710)	(1.207;1.564)	(1.587;1.902)	(1.725;1.865)	(1.405;1.946)	(1.455;1.085)	(1.990;1.990)	(1.450;1.100)
France	2010	(1.094;1.585)	(1.209;1.370)	(1.271;0.596)	(1.592;0.559)	(1.326;0.787)	(1.136;1.810)	(1.705;1.199)	(1.145;1.349)
Germany	2008	(1.092;1.264)	(1.582;0.801)	(1.921;1.960)	(1.386;1.349)	(1.211;1.919)	(1.093;0.506)	(1.418;1.334)	(1.111;1.537)
Germany	2009	(1.206;1.548)	(1.459;1.944)	(1.599;1.440)	(1.078;1.042)	(1.926;1.939)	(1.670;1.924)	(1.995;1.534)	(1.634;1.294)
Germany	2010	(1.079;1.478)	(1.740;1.019)	(1.445;1.997)	(1.070;1.467)	(1.051;1.402)	(1.355;0.513)	(1.052;1.465)	(1.755;1.409)
Italy	2008	(1.527;0.700)	(1.888;0.818)	(1.390;1.270)	(1.918;1.658)	(1.492;1.968)	(1.316;0.571)	(1.880;1.688)	(1.861;0.730)
Italy	2009	(1.424;1.635)	(1.258;1.860)	(1.410;1.537)	(1.768;0.540)	(1.749;1.010)	(1.602;1.523)	(1.820;1.357)	(1.584;0.981)
Italy	2010	(1.208;1.349)	(1.753;0.700)	(1.054;0.573)	(1.239;1.627)	(1.081;0.638)	(1.916;1.075)	(1.832;1.962)	(1.579;1.920)
Netherlands	2008	(1.477;0.600)	(1.474;1.212)	(1.347;0.999)	(1.327;1.563)	(1.654;1.881)	(1.498;1.206)	(1.803;1.194)	(1.933;0.800)
Netherlands	2009	(1.442;1.494)	(1.789;0.933)	(1.451;0.652)	(1.773;0.767)	(1.361;0.671)	(1.648;1.571)	(1.073;1.275)	(1.206;1.293)
Netherlands	2010	(1.671;0.856)	(1.693;1.854)	(1.524;1.373)	(1.848;0.762)	(1.995;0.642)	(1.764;1.611)	(1.587;1.547)	(1.973;1.935)
Portugal	2008	(1.905;0.788)	(1.012;1.783)	(1.921;0.881)	(1.927;1.345)	(1.892;0.579)	(1.846;0.650)	(1.744;1.854)	(1.705;1.630)
Portugal	2009	(1.774;1.870)	(1.169;1.768)	(1.405;1.738)	(1.264;0.723)	(1.351;1.919)	(1.650;1.040)	(1.984;1.274)	(1.538;1.953)
Portugal	2010	(1.184;1.865)	(1.855;1.444)	(1.002;1.628)	(1.663;1.721)	(1.696;1.562)	(1.463;0.742)	(1.869;1.664)	(1.462;1.166)

Extensions

- Two important extensions:

1) *Compositional IO analysis* – the cells become „unordered/categorical bin charts“

- In terms of Leontief formulas, basic CoDa operations like $\oplus$, $\ominus$, $\odot$, $\boxdot$ can be used

2) *Functional IO analysis* – the cells are functions of „intermediary production“ – the main value of the cell

- Largely, this depends on the nature of the (intervening) variable we condition upon, e.g. size of companies (nr. of employees, revenues, etc.), their sociodemographics or other features
- To be done in future work

Conclusion – scientific relevance

- Derivation of a *completely new way* of approaching stochastic possibilities of input-output analysis
- Note: everything is done under the assumption of independent and identically distributed random variables
- Significant gain in information, the gain in accuracy and predictability still to be tested
- Questions: computing and theoretical issues (both: inverting a square matrix of random variables)
- Calculus of distributions, that could form also the foundation of the work in symbolic data analysis and the analysis of complex data (also FDA and CoDA) in future, where the work so far has largely been limited to general uniform distribution assumptions

THANK YOU FOR LISTENING!

Q&A

andrej.srakar@ier.si