**WP. 14**
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE**
**CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION**
**STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)**

**Joint UNECE/Eurostat work session on statistical data confidentiality**
(Bilbao, Spain, 2-4 December 2009)

Topic (ii): Synthetic and hybrid data

# ARTIFICIAL DATA THROUGH CALIBRATION AND EMPIRICAL COPULAS

**Invited Paper**

Prepared by Flavio Foschi (ISTAT, Italy)

# Artificial data through calibration and empirical copulas

Flavio Foschi*

Istat, Division for Information Technology and Methodology, via Cesare Balbo 16, 00184 Rome, Italy, foschi@istat.it

**Abstract:** The aim of this study is to investigate artificial data generation possibilities offered by calibration estimators and empirical copulas, for continuous or mixed support variables with high skewness and kurtosis. After a brief discussion of the method, results of exercises on Survey Household Income data released by the Bank of Italy are shown and a global protection indicator is proposed.

## 1 Introduction

Simulation of multivariate economic data is a challenging task, from a methodological and computational point of view. Concerning the former, models lacking in appropriate constraints generally suffer of inadequate ability to fit features as high skewness and kurtosis. For the latter, moment based models (like maximum entropy ones) request a computational effort growing more than linearly with the number of variables. Difficulties due to the curse of dimensionality problem are obvious, but also the discretization of a wide univariate support can be troublesome: the simplification of circumventing the resort to integration is paid with the necessity of a fine grid resolution. Moreover for variables having probability masses concentrated to discrete points of the support many computational resources could be saved focusing only on support values relevant to efficiently describe examined phenomena. Disregarding "true" Data Generating Process (DGP) oriented investigations, a proposal to cope with these problems retaining main observed data features is presented in section two; simulation studies performed on Survey Household Income data released by the Bank of Italy are described in section three and a global protection measure is proposed; section four contains a summary and outlook on future research.

## 2 Artificial data generation

The merit of the simulation method would be the possibility of avoiding parametric assumptions on the DGP in a multivariate framework, maintaining some dependence relationships between variables (and eventually between observations, for time series or panel data). Subsections 2.1 and 2.2 discuss this matter.

* The author thanks Dr. Luisa Franconi and Dr. Daniela Ichim for useful suggestion and comments.

### 2.1 Empirical likelihoods and copulas

The intention of simulating multivariate economic data efficiently, from a computational and statistical point of view, can be pursued using calibration weights and empirical copulas. The first are instrumental to the reduction of the univariate discretized support dimension for each density to approximate, while the second is necessary to recovery a joint probability law from the univariate ones. An interesting set of calibration weights is constituted by those which maximize the empirical likelihood (Owen, 2001). The matter is discussed in several works (i.e. C. Wu 2004). Although such estimates are first order equivalent to those obtained by linear regression (that is minimizing euclidean squared distances between original and constrained weigths), avoiding their explicit dependence from the cross-product data matrix positively affects the robustness against influential observations. Given constraints which can be put in the form of orthogonality conditions (let $\boldsymbol{\psi}_i$ be the row-vector of conditions for the $i^{\text{th}}$ observation, $i$=1,…,$n$), weigths $w_i$ are achieved minimizing the Kullback-Leibler divergence w.r.t those assigned to each statistical unit in the sample. Within the simple random sampling scheme, the Empirical Maximum Likelihood $\boldsymbol{w}$ minimizes $D_{KL}=\Sigma(1/n)ln[(1/n)/w_i]$. The Lagrangean is

$$L = -\sum \frac{1}{n} ln\left(nw_i\right) - \gamma\left(\sum w_i - 1\right) - \sum w_i \boldsymbol{\lambda}^t \boldsymbol{\psi}_i$$

First order condition equations are:

$$\partial L/\partial w_i \equiv -\frac{1}{nw_i} - \gamma - \boldsymbol{\lambda}^t \boldsymbol{\psi}_i = 0; \qquad \partial L/\partial \gamma \equiv \left(1 - \sum w_i\right) = 0; \qquad \partial L/\partial \boldsymbol{\lambda} \equiv \sum w_i \boldsymbol{\psi}_i^t = 0$$

From the first, it results:

$$w_i = \frac{1}{n} \frac{1}{-\gamma - \boldsymbol{\lambda}^t \boldsymbol{\psi}_i}$$

Since $\Sigma[-1/nw_i - \gamma - \boldsymbol{\lambda}^t \boldsymbol{\psi}_i]\ w_i = 0$, due to constraints, $-\gamma$=1 follows. As we will see later, in the proposed framework EML weigths are used to calibrate a subset of the original sample to fit several moment conditions.

The study of copula functions has received growing attention in recent years and a wide review is presented in Nelsen (2006). The Sklar's theorem (Rueschendorf, 2009), establishes that given an $m$-dimensional distribution function $F$, with marginals $F_1$, $F_2$, …, $F_m$, there exists an $m$-dimensional distribution function $C$ (the copula function) on $[0,1]^m$ with uniform marginals, such that

$$F\left(x_1, x_2, \cdots, x_m\right) = C\left[F_1\left(x_1\right), \cdots, F_m\left(x_m\right)\right]$$

For a continuous random variable $X$, $F(X) \equiv p(X < x)$ and posing $U \equiv U(0,1)$ the distributional transform can be defined as $X=F^{-1}(U)$; then

$$p\left[\bigcap\left(X_j \le x_j\right)\right] = p\left\{\bigcap\left[F_j^{-1}\left(U_j\right) \le x_j\right]\right\} = p\left\{\bigcap\left[U_j \le F_j\left(x_j\right)\right]\right\} \equiv C\left[F_1\left(x_1\right), \cdots, F_m\left(x_m\right)\right]$$

If marginals $F_j$ are not known, their sample counterparts, empirical cumulative distribution functions, can be used and the inverse distributional transform delivers vectors proportional to observed ranks:

$$n\hat{U}_j = n\hat{F}_j\left(X_j\right) = R_j \qquad \left(j = 1, \cdots, m\right)$$

Hence, the Empirical Copula Function is

$$\hat{C}\left(\mathbf{t}\right) = \frac{1}{n}\sum I\left[\bigcap\left(\hat{U}_j \le t_j\right)\right] = \frac{1}{n}\sum I\left[\bigcap\left(R_j \le nt_j\right)\right] \qquad \mathbf{t} \in \left[0,1\right]^m$$

This circumstance clarifies the key role of the observed rank matrix in conveying dependence relationships featuring data. If $\{x_{(1)j}, x_{(2)j}, \ldots, x_{(n)j}\}$ and $\{r_{1j}, r_{2j}, \ldots, r_{nj}\}$ are respectively order statistics and ranks for the $j^{th}$ variable, the dependence structure preserving is (approximately) accomplished posing

$$x_{ij} \equiv x_{(r_{ij})j} \qquad \left(i = 1, \cdots, n \qquad j = 1, \cdots, m\right)$$

Hence, no rank swapping is performed: ECF does not play any role in data perturbation.

### 2.2 A synthesis of the procedure

A sequence of steps summarizes the proposed data generation method:

a) for each variable, by means of $k$–1 empirical percentiles, $k$ strata of the observed support are delimited;

b) the observed range is enlarged by values obtained adding to the original maximum (minimum) a positive (negative) term related to data sparsity; that step is needed to ensure the existence of a solution for the constrained optimization problem which defines calibration weights;

c) for each stratum, $h$ values are drawn from a uniform law; hence a discretized version of each variable support is defined. $kh$ is the number of points of the discretized support; that step allows, together with step (a), to control the trade-off between accuracy w.r.t. distributional features and data protection; calibration is needed to make the distribution a satisfactory approximation of the density which generates data;

d) moments up to order $p$ are used to calibrate weights for the artificial discrete support of length $kh$ generated in (c); so doing the information loss due to the support compression of steps (a) and (c) is partially corrected; it is necessary to stress that calibration weights are only used to draw samples from the compressed support, satisfying moment constraints;

e) according to calibration weights, samples of length $n$ (with repetition if $n > kh$) are drawn;

f) each sample from (e) is ordered according to observed ranks; that step is performed one variable at a time, maintains univariate dependence relationships between statistical units and implicitly recovers multivariate links between variables included in the data set; no rank swapping is performed;

This procedure is somehow similar to the maximum entropy bootstrap (Vinod, 2004), implicitly embedding a clustering task, points (a) and (c), and explicitly imposing moment constraints, point (d). Moment based estimations of probability laws are discussed in several papers (i.e. X. Wu 2003). In our framework, aims other than observed features replication are not considered, omitting any inferential purpose about "true" probability laws. At least first four moments are included, introducing further constraints (that is higher moment conditions) until the null hypothesis of equal distributions is not refused.

# 3 Simulated data for the survey of household income

## 3.1 Original data and simulation design

Data gathered by the 2006 Survey on Household Income and Wealth, freely available by the Bank of Italy Economic and Financial Statistics Department (see http://www.bancaditalia.it/statistiche/indcamp/bilfait/dismicro), seem interesting to assess the fitting accuracy for extremely sparse variables. The absence of socio-economic interpretative intentions leads to consider a subset of variables, resulting by aggregations based on income origins; grossing up weights are ignored. Data, in euro currency units, contains 7768 records and five variables: *Ycf* (income from financial assets), *Yl* (payroll income), *Yt* (pensions and net transfers), *Ym* (net self-employment income), *Yca* (income from real-estate). Extreme degrees of data compression are considered for investigating the trade-off between accuracy and protection. Relevant parameters are the number of support slices and sample fractions. Two settings (3 strata with 20% of sampled units and 100 strata with 50% of sampled units) are separately considered in 2500 replications, repeating steps (e)-(f) 50 times for each outcome of steps (a)-(d).

| | *Ycf* | *Yl* | *Yt* | *Ym* | *Yca* |
|---|---|---|---|---|---|
| *Min* | -25432.36 | 0.00 | 0.00 | -20000.00 | 0.00 |
| *p25* | 3.42 | 0.00 | 0.00 | 0.00 | 2400.00 |
| *p50* | 67.25 | 6105.00 | 6518.00 | 0.00 | 6000.00 |
| *p75* | 298.20 | 20000.00 | 14690.00 | 0.00 | 8400.00 |
| *Max* | 99789.76 | 251000.00 | 429770.00 | 800000.00 | 152000.00 |

**Table 3.1** Some order statistics for survey data.

| | *Ycf* | *Yl* | *Yt* | *Ym* | *Yca* |
|---|---|---|---|---|---|
| *Mean* | 280.54 | 12054.03 | 8781.26 | 4327.05 | 6564.37 |
| *Std. Dev.* | 2742.24 | 15355.02 | 11194.32 | 19483.57 | 7457.06 |
| *Skewness* | 11.44 | 2.11 | 7.88 | 18.21 | 5.66 |
| *Kurtosis* | 322.74 | 17.07 | 261.62 | 549.23 | 72.47 |

**Table 3.2** First four cumulants for survey data

| | *Ycf* | *Yl* | *Yt* | *Ym* | *Yca* |
|---|---|---|---|---|---|
| *Ycf* | 1.00 | -0.02 | 0.15 | 0.12 | 0.25 |
| *Yl* | -0.02 | 1.00 | -0.37 | -0.07 | 0.11 |
| *Yt* | 0.15 | -0.37 | 1.00 | -0.09 | 0.16 |
| *Ym* | 0.12 | -0.07 | -0.09 | 1.00 | 0.24 |
| *Yca* | 0.25 | 0.11 | 0.16 | 0.24 | 1.00 |

**Table 3.3** Correlation matrix for survey data

### 3.1.1 Accuracy

Focusing on the univariate field, survey data peculiarities appear to be satisfactory taken by statistical units simulated with the second setting (100 strata, 50% of sampled units). Since the calibration is only performed on moment conditions because of the simple implementation of constraints, as a by-product of the cumulants reproduction ability, similarity between observed and simulated order statistics is achieved. By comparison of tables 3.1 and 3.4, the first setting delivers less accurate results than the second and simulated ranges are generally reduced; for *Ym* in the first setting a strong shrinking of the maximum can be noticed; the overshooting for the minimum of *Ym* is instrumental to the approximation (from lower values) of skewness and kurtosis; for *Yl*, the median and the 3$^{rd}$ quartile are oppositely distorted. In the second setting the maximum of *Ym* is better reproduced and the overshooting for the minimum is softened. In the second setting order statistics are generally closer to those observed; confidence intervals (CIs) referred to *Yca* quartiles do not comprise those observed and the same happens, with lower intensity, for the minimum of *Ym*; the CI for the 3$^{rd}$ quartile of *Yl* weakly exceeds the sample estimation as well values simulated for 2$^{nd}$ and 3$^{rd}$ quartile of *Ycf*. *Ym* is the extreme paradigm of the protection-accuracy conflict: for a variable whose tails contain all relevant information for describing the distribution, all intensities are fully involved by the distortion due to the protection. A better behaviour can be noticed about cumulants (tables 3.2 and 3.5). In the first setting, *Ym* skewness and kurtosis are downwardly approximated; adjusting skewness and kurtosis leads to underestimate the mean value of *Yt*; *Ycf* markedly underestimates 3$^{rd}$ and 4$^{th}$ cumulants. In the second setting simulated quantities are closer to those observed; a weak underestimation for the *Yca* mean value is noticed but CI are shorter. Better results could be achieved in the first setting tuning the number of moment constraints for the variables whose density approximation should be improved; such exercise has not been performed the focus

being on methods suitable for a large number of variables, as this is more relevant for statistical institutes.

| | | 3 strata, 20% sampled units | | | | | 100 strata, 50% sampled units | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *Min* | *p25* | *p50* | *p75* | *Max* | *Min* | *p25* | *p50* | *p75* | *Max* |
| *Ycf* | *p2.5* | -20238.3 | 0.0 | 48.1 | 108.4 | 21686.5 | -29517.5 | 0.0 | 69.2 | 313.5 | 43879.0 |
| | *Mean* | -16460.8 | 0.0 | 50.9 | 109.7 | 25547.8 | -25814.2 | 2.4 | 74.2 | 333.5 | 89611.9 |
| | *p97.5* | -13423.6 | 0.0 | 53.6 | 111.8 | 32111.3 | -19539.3 | 6.1 | 77.9 | 360.0 | 112583.3 |
| *Yl* | *p2.5* | 0.0 | 0.0 | 9092.2 | 14245.9 | 79349.2 | 0.0 | 0.0 | 4514.6 | 20425.8 | 119573.4 |
| | *Mean* | 0.0 | 0.0 | 9479.2 | 14463.7 | 84551.9 | 0.0 | 0.0 | 7082.0 | 21146.1 | 219428.3 |
| | *p97.5* | 0.0 | 0.0 | 9608.0 | 14722.1 | 97549.2 | 0.0 | 0.0 | 8305.3 | 21802.5 | 278502.7 |
| *Yt* | *p2.5* | 0.0 | 0.0 | 3568.8 | 9419.1 | 89644.0 | 0.0 | 0.0 | 6503.1 | 14633.9 | 107957.3 |
| | *Mean* | 0.0 | 0.0 | 3674.6 | 9431.8 | 136548.4 | 0.0 | 0.0 | 6653.3 | 15130.7 | 314007.7 |
| | *p97.5* | 0.0 | 0.0 | 3979.0 | 9528.1 | 643438.6 | 0.0 | 0.0 | 7082.6 | 15526.5 | 503211.3 |
| *Ym* | *p2.5* | -29171.0 | 0.0 | 0.0 | 0.0 | 394210.3 | -23823.2 | 0.0 | 0.0 | 0.0 | 388214.7 |
| | *Mean* | -29170.5 | 0.0 | 0.0 | 0.0 | 591466.1 | -23823.2 | 0.0 | 0.0 | 0.0 | 679318.0 |
| | *p97.5* | -29171.0 | 0.0 | 0.0 | 0.0 | 622144.9 | -23823.2 | 0.0 | 0.0 | 0.0 | 923022.9 |
| *Yca* | *p2.5* | 308.3 | 2389.1 | 5250.4 | 6538.0 | 79316.5 | 0.0 | 0.0 | 5387.2 | 8580.8 | 115848.5 |
| | *Mean* | 629.4 | 2684.9 | 5272.3 | 6551.8 | 142172.2 | 0.0 | 0.0 | 5596.2 | 8756.4 | 146122.7 |
| | *p97.5* | 777.8 | 2901.2 | 5277.7 | 6601.6 | 160687.3 | 0.0 | 0.0 | 5865.9 | 9122.0 | 162847.4 |

**Table 3.4** Order statistics summary

| | | 3 strata, 20% sampled units | | | | 100 strata, 50% sampled units | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *Mean* | *Std. Dev.* | *Skewness* | *Kurtosis* | *Mean* | *Std. Dev.* | *Skewness* | *Kurtosis* |
| *Ycf* | *p2.5* | 112.7 | 2409.5 | 1.8 | 16.6 | 252.2 | 2041.7 | 5.8 | 126.7 |
| | *Mean* | 167.3 | 2537.5 | 2.3 | 19.4 | 309.7 | 2614.1 | 13.3 | 389.9 |
| | *p97.5* | 223.3 | 2663.5 | 2.7 | 22.8 | 369.5 | 3265.0 | 19.9 | 691.7 |
| *Yl* | *p2.5* | 11303.2 | 14757.8 | 1.4 | 1.1 | 11836.9 | 14641.9 | 1.5 | 4.3 |
| | *Mean* | 11562.4 | 15070.7 | 1.4 | 1.4 | 12150.7 | 15269.1 | 2.1 | 13.8 |
| | *p97.5* | 11860.2 | 15424.8 | 1.5 | 1.7 | 12479.2 | 16006.4 | 3.1 | 30.6 |
| *Yt* | *p2.5* | 7177.2 | 11097.6 | 2.4 | 6.3 | 8574.9 | 9830.9 | 1.6 | 8.3 |
| | *Mean* | 7431.8 | 11635.8 | 3.3 | 45.8 | 8813.1 | 11078.5 | 6.6 | 187.8 |
| | *p97.5* | 7694.0 | 13634.5 | 14.9 | 633.6 | 9065.2 | 13349.1 | 15.4 | 531.6 |
| *Ym* | *p2.5* | 3882.5 | 14953.7 | 6.7 | 96.3 | 4190.8 | 14490.9 | 11.4 | 237.4 |
| | *Mean* | 4287.1 | 18892.0 | 13.7 | 327.7 | 4583.6 | 18848.6 | 17.1 | 480.4 |
| | *p97.5* | 4729.9 | 23116.8 | 17.9 | 498.6 | 5018.3 | 24351.6 | 23.4 | 909.1 |
| *Yca* | *p2.5* | 6097.0 | 6806.6 | 2.8 | 9.3 | 5855.3 | 7147.9 | 3.5 | 30.9 |
| | *Mean* | 6372.9 | 7377.6 | 4.5 | 47.6 | 6040.8 | 7805.7 | 5.1 | 59.1 |
| | *p97.5* | 6585.1 | 8011.4 | 6.1 | 79.5 | 6215.4 | 8551.7 | 6.5 | 86.3 |

**Table 3.5** Cumulants summary

Comparing tables 3.3 and 3.6 pairwise correlations are satisfactory close, an important circumstance for a multivariate validation of the simulation framework. In both simulation settings, correlation signs are maintained; in the first one, magnitudes are closer to those measured on data, but correlations often exceed the simulated CI

boundaries. However, in both settings CI lengths look short and, with the exception of that related to (*Yl*, *Yt*) in the second setting (length of 0.11), all others are not greater than 0.06.

| | 3 strata, 20% units sampled | | | 100 strata, 50% units sampled | | |
|---|---|---|---|---|---|---|
| | *p2.5* | *Mean* | *p97.5* | *p2.5* | *Mean* | *p97.5* |
| *Ycf, Yl* | -0.06 | -0.05 | -0.04 | -0.03 | -0.01 | 0.00 |
| *Ycf, Yt* | 0.18 | 0.21 | 0.22 | 0.11 | 0.14 | 0.17 |
| *Ycf, Ym* | 0.07 | 0.10 | 0.12 | 0.09 | 0.12 | 0.15 |
| *Ycf, Yca* | 0.19 | 0.21 | 0.23 | 0.18 | 0.21 | 0.24 |
| *Yl, Yt* | -0.31 | -0.30 | -0.26 | -0.42 | -0.38 | -0.31 |
| *Yl, Ym* | -0.09 | -0.08 | -0.07 | -0.09 | -0.07 | -0.06 |
| *Yl, Yca* | 0.11 | 0.12 | 0.13 | 0.11 | 0.12 | 0.13 |
| *Yt, Ym* | -0.09 | -0.07 | -0.06 | -0.13 | -0.10 | -0.07 |
| *Yt, Yca* | 0.19 | 0.21 | 0.22 | 0.14 | 0.15 | 0.17 |
| *Ym, Yca* | 0.19 | 0.22 | 0.24 | 0.20 | 0.23 | 0.26 |

**Table 3.6** Correlation coefficients summary

Hints about the retention ability for more complex dependence relationships can be gathered from hypothesis tests as:

$$H_0 : \frac{\hat{m}_{a,b,c,d,e}}{m_{a,b,c,d,e}} - 1 = 0, \quad m_{a,b,c,d,e} \equiv \left( \frac{1}{n} \sum Ycf_i^{\,a} \cdot Yl_i^{\,b} \cdot Yt_i^{\,c} \cdot Ym_i^{\,d} \cdot Yca_i^{\,e} \right)^{\frac{1}{a+b+c+d+e}}$$

In details, 500 mixed moments have been selected raising each variable to a power given by a random integer in [0, 4]. Hypothesis tests are intended to assess if simulated mixed moments are significantly different from original ones; those tests are based on CIs estimated by means of 2,500 replications of data simulation. The Mean Absolute Percentage Error (MAPE) is calculated for each mixed moment.

| | Order | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Frequency | 3 | 9 | 16 | 34 | 47 | 48 | 58 | 59 | 61 | 61 | 40 | 31 | 18 | 8 | 6 | 1 |
| 3 strata, 20% s.u. | $H_0$ not false | 0.33 | 0.56 | 0.56 | 0.50 | 0.57 | 0.58 | 0.59 | 0.64 | 0.69 | 0.57 | 0.55 | 0.45 | 0.67 | 0.38 | 0.67 | 0.00 |
| | *MAPE* | 0.10 | 0.20 | 0.19 | 0.15 | 0.17 | 0.15 | 0.14 | 0.13 | 0.10 | 0.11 | 0.07 | 0.07 | 0.05 | 0.04 | 0.04 | 0.09 |
| 100 strata, 50% s.u. | $H_0$ not false | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.98 | 0.98 | 0.98 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | *MAPE* | 0.04 | 0.10 | 0.10 | 0.12 | 0.12 | 0.13 | 0.14 | 0.13 | 0.11 | 0.10 | 0.10 | 0.08 | 0.09 | 0.07 | 0.08 | 0.08 |

**Table 3.7** Proportions of null not refused and MAPEs w.r.t. moment orders

Table 3.7 shows percentage of instances in which the null is not refused at the level $\alpha$=0.05 and MAPEs, both grouped w.r.t. moment orders. The frequency about the hypothesis tests related to a certain order is calculated as the ratio between the number of instances in which 0 falls within CI boundaries and the number of mixed moments having that order. Analogously, MAPEs referred to a certain order are

averaged. In the first setting, the dispersion around test statistics is only partially accounted for, while averaged MAPEs are greater than those calculated in the second setting up to the 8th order. Beyond the order 8, lower support compressions do not improve point estimations: in both settings supports discretization leads to a partially unsatisfactory description of mixed moment domains. Hence, simulated data can preserve linear relationships between pairs of variables as well fairly high mixed moments. Figure 3.1 depictes simulated CIs for a sample of mixed moments randomly drawn from those tested in the second setting (in x-axis exponents are shown, respectively referred to *Ycf, Yl, Yt, Ym, Yca*).

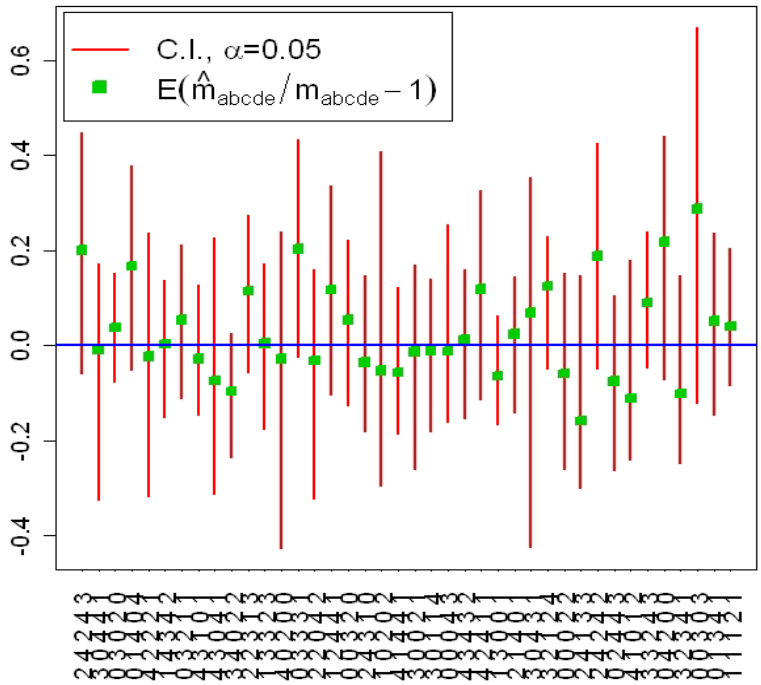


**Fig 3.1** Some mixed moment tests (100 strata, 50% sampled units)

Concluding accuracy analyses, a test which provides a general measure of equality for multivariate distributions is suitable; that proposed by Székely and Rizzo† (2004) has been used for each simulated sample. Under the null, a random permutation of pooled original and simulated observations is equal in distribution to a random sample drawn from the mixture of sources. Due to data size, in each simulation, 100 tests on pooled random subsamples (100 observations from each source) have been performed; overall averaged p-values are 0.328 for the first setting and 0.999 for the second.

† For two samples **x** and **y** having respectively size $n_1$ and $n_2$ the test statistic is:

$$\varepsilon_{n_1,n_2} \equiv \left(\frac{n_1 n_2}{n_1+n_2}\right) n_1^{-2} n_2^{-2} \sum_i \sum_j \sum_l \sum_h \left\|x_i - y_l\right\| + \left\|x_j - y_m\right\| - \left\|x_i - x_j\right\| - \left\|y_l - y_m\right\|$$

### 3.1.2 A global protection indicator for simulated data

A proposal for quantifying the global protection implied by simulated data can be scheduled as follows:

a) the original dataset (with observations $x_{ij}$, $i$=1,…,$n$ $j$=1,…,$m$) is overfitted choosing the highest number of clusters between those admissible according to a selection criterion; let $L$ be the set of clusters with size 1 having elements $l_s$ ,

b) each simulated record ($y_{ij,}$ $i$=1,…,$n$ $j$=1,…,$m$) is assigned to the closest centroid identified in (a); let $H$ be the set of non-empty clusters having centroids in $L$ and $H_{ls}$ the cluster around $l_s$,

c) for each $l_s \in L$, if $l_s$ is the $s^{th}$ element of $L$, $D_g \equiv max_j|x_{sj}-median(x_j)|$ is retained (that is, the worst dimension in term of identifiable units is considered),

d) for each $l_s \in L$, if $\boldsymbol{y}_u$ is the $u^{th}$ record of $H_{ls}$, $d_s \equiv min_u|x_{sg}-y_{ug}|/D_g$ is calculated and a threshold $q \in [0,1]$ is defined,

e) $ns \equiv |H| / |L|$ expresses the proportion of simulated records potentially not secured; $ans \equiv ns \cdot E_s[I(d_s < q)]$ adjusts $ns$ for the proportion of records subjectively secured because of the proximity of simulated and median values for the less favourable variable detected in (c),

f) the global protection indicator is defined as $gp \equiv 1-ans$.

About points (c) and (d), alternative ways for associating a unique measure to each record could be obtained choosing some syntheses of gaps; standardizations different from the proposed one can be applied but it is to notice that the interquartile range is not always useful to this aim due to its equality to zero for some variables; other possibilities are related to standardization of variables or rotation of Cartesian axes. For simplicity, the first step has been implemented resorting to the *k-means* algorithm, choosing the highest number of clusters, 3981, which locally maximizes the Calinski-Harabasz *pseudo-F* (Sugar and James, 2003) and minimizes the averaged *within* sum of squares; through overfitting, the lack of detection for simulated units far from the region where the multivariate density is predominantly allocated should be avoided. From table 3.8, if $q$ = 0.1, roughly 40% of simulated records belonging to $H$ are considered critical in the first setting and the proposed protection measure can be calculated as 1–0.4·0.1185=0.95; in the second setting, $q$ is less than 0.1 for the 90% of instances and $gp$ is 1–0.9·0.293=0.74. Such behaviours are outcomes of (different) simulation abilities to mimic distribution details, induced by data generation settings.

| 3 strata, 20% s.u. | $ns$ | $q$ | 0.00 | 0.03 | 0.08 | 0.14 | 0.21 | 0.28 | 0.35 | 0.49 | 0.60 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.1185 | $E_s[I(d_s < q)]$ | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 |
| 100 strata, 50% s.u. | $ns$ | $q$ | 0.00 | 0.01 | 0.01 | 0.02 | 0.03 | 0.04 | 0.06 | 0.08 | 0.11 |
| | 0.2930 | $E_s[I(d_s < q)]$ | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 |

**Table 3.8** Elements for estimating the global protection indicator

## 4 Some considerations

Within a sampling setting, a data simulation method has been proposed. The trade-off between accuracy and protection against disclosure is controlled through the compression of univariate supports. Calibration weights obtained by means of moment conditions partially correct the information loss, so that samples drawn from artificial supports approximately retain univariate statistical features. Multivariate dependence relationships are maintained ordering simulated samples according to the matrix of ranks. The irrelevance of explicit conditional independence assumptions regarding subsets of variables provides robustness against misspecifications. Moreover, compression of supports and empirical copula expression of multivariate dependence jointly reduce the computational burden: while the former decreases the number of support points, the latter ensures a growth only linear w.r.t. the number of variables. Exercises on Survey Household Income data collected by the Bank of Italy, have shown encouraging results. A global protection indicator has been proposed unifying two aspects: detection of statistical units remote to the multivariate density core and evaluation of the exchangeability between simulated ones. Extreme settings for data generation let understand the existence of tuning opportunities for accuracy and protection preferences. It remains to assess the behaviour of non linear combinations of simulated data when records are grouped according to known homogeneity features: further capability studies seem necessary.

**Acknowledgments**. Istat is not responsible for any view or result presented. The author was supported by the European Project ESSnet-SDC in the field of Statistical Disclosure Control.

## References

Nelsen, R.B. (2006). *An Introduction to Copula.* Springer-Verlag.

Owen, A.B. (2001). *Empirical Likelihood.* Chapman & Hall.

Rueschendorf, L. (2009). On the distributional transform, Sklar's theorem, and the empirical copula process. *Journal of Statistical Planning and Inference.* 139(11). 3921-3927.

Sugar, C., James, G. (2003) Finding the Number of Clusters in a Data Set: An Information Theoretic Approach. *Journal of the American Statistical Association.* 98. 750-763.

Székely, G. J., Rizzo, M. L. (2004). Testing for Equal Distributions in High Dimension. *InterStat*. Nov. (5).

Vinod, H. (2004). Ranking mutual funds using unconventional utility theory and stochastic dominance. *Journal of Empirical Finance.* 11(3). 353-377.

Wu, C. (2004). Weighted Empirical Likelihood Inference. *Statistics and Probability Letters*. 66(1). 67-79.

Wu, X. (2003). Calculation of maximum entropy densities with application to income distribution. *Journal of Econometrics*. 115(2). 347-354.