

ESSnet Meeting, Rome November 3-4, 2009

Artificial Data Generation: a Proposal

Flavio Foschi
(foschi@istat.it)

Summary

1. Main characteristics of the method
2. Main steps of the simulation procedure
3. Test : the survey on household income
 - a. Information loss
 - i. Order statistics
 - ii. Cumulants
 - iii. Correlations
 - iv. Mixed moments
 - b. A global risk indicator
4. Conclusions and further work

1. Main characteristics

- **Model free method:** robust method
 - no dependence nor independence hypothesis are needed;
 - no parametric modelling;
- **Semplicity:** the simulation from multivariate distributions is treated as sum of univariate probabilities => computational complexity grows linearly with the number of dimensions

1. Main characteristics

- **Sound theoretical base:**
 - calibration of weights by means of moment conditions accommodates information loss;
 - the approach of empirical copulas (Sklar theorem) maintains multivariate dependence relationships simply by using univariate characteristics;
- **Great flexibility:** tuning possibilities for accuracy and protection are allowed by selecting:
 - the number of strata in which the support is divided
 - the number of observations generated.

1. Main characteristics

- **Global disclosure risk indicator** is defined by both:
 - the identification of statistical units remote from the highest density region;
 - evaluation of the exchangeability of simulated observations.

2. Main steps of the simulation procedure

- a) for each variable, by means of $k-1$ empirical percentiles, k strata of the observed support are delimited;
- b) for each stratum, h values are drawn from a uniform law;
- c) moments up to order p are used to calibrate weights for the artificial discrete support of length kh generated in (b);
- d) according to calibration weights, samples of length n are drawn;
- e) each sample from (d) is ordered according to the observed ranks (empirical copulas).

3. Test: Survey on Household Income 2006

SHIW data (<http://www.bancaditalia.it/statistiche/indcamp/bilfait/dismicro>), seem interesting to assess the fitting accuracy for extremely sparse variables. A subset of them is considered:

- Y_{cf} : income from financial assets,
- Y_l : payroll income,
- Y_t : pensions and net transfers,
- Y_m : net self-employment income,
- Y_{ca} : income from real-estate.

Opposite settings (3 strata with 20% of sampled units and 100 strata with 50% of sampled units), are separately considered in 2500 replications, repeating steps (d)-(e) 50 times for each outcome of steps (a)-(c).

3a. Order Statistics

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

Table 1 Some order statistics for survey data.

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

Table 2 Order statistics summary for simulated data

3b. Cumulants

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

Table 3 First four cumulants for survey data

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

Table 4 Cumulants summary

3c. Correlations

	<i>Ycf</i>	<i>Yl</i>	<i>Yt</i>	<i>Ym</i>	<i>Yca</i>
<i>Ycf</i>	1.00	-0.02	0.15	0.12	0.25
<i>Yl</i>	-0.02	1.00	-0.37	-0.07	0.11
<i>Yt</i>	0.15	-0.37	1.00	-0.09	0.16
<i>Ym</i>	0.12	-0.07	-0.09	1.00	0.24
<i>Yca</i>	0.25	0.11	0.16	0.24	1.00

Table 5 Correlation matrix for survey data

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

Table 6 Correlation coefficients summary

3d. Mixed Moments

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

		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 7 Proportion of null not refused and MAPE w.r.t. moment orders

Each CI ($\alpha=0.05$) is obtained from the outcomes of 2,500 replications. Proportions of those for which H_0 is not false are shown in table 7 according to respective moment orders. Mean Absolute Percentage Errors are similarly averaged.

4. A Global Risk Indicator

- In a multivariate setting it is difficult to define observations at risk;
- We analyse the dimension which is worst in terms of outliers i.e. identifiable units;
- To identify outliers use clustering methods: because of the large computational burden the method use is aggregational and not hierarchical;
- In order to find all possible outliers we overfit the data;
- Once outliers have been identified we simulate the data;
- We match original and simulated data and define a measure of the deviation from original data that varies from 0 (no risk) to 1 (risk);
- Set a threshold q to define when an observation is at risk;
- Find the proportion of simulated records potentially at risk r_p ;
- Adjust r_p for the less favourable variable detected.

4a. Example

The step (a) has been implemented resorting to the *k-means* algorithm. Through overfitting, lack of detection for simulated units far from the multivariate density core should be avoided.

3 strata, 20% s.u.	<i>rp</i>	<i>q</i>	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.	<i>rp</i>	<i>q</i>	0.00	0.01	0.01	0.02	0.03	0.04	0.06	0.08	0.11
	0.2930	$E_s[I(d_s \leq q)]$	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90

Table 8 Elements for estimating the adjusted risk proportion indicator

I.e., let be $q = 0.1$. Then, w.r.t. the first setting, roughly 40% of simulated records which mimic isolated true observations are considered critical and the proposed global risk measure decreases to 0.05; in the second setting, q is less to 0.1 for 90% of instances and the global risk measure reduces only to 0.26.

5. Conclusions and further work

- **Very encouraging results;**
- **The simplicity of the method allows the application to quite complex cases;**
- **Currently we are running experiments on enterprise data.**