

A Heuristic Criterion to Rank Statistical Units in the Multivariate Analysis

Un Criterio Euristico per l'Ordinamento delle Unità Statistiche nell'Analisi Multivariata

Egidio Cascini

Studio Esercizi di Qualità, via Mosè Bianchi, 103 – 20149 Milano
ecascini@mmm.com

Riassunto: Il lavoro propone un criterio per l'ordinamento di unità statistiche multivariate, mediante un indicatore univariato, variabile nell'intervallo 0,1. Viene mostrato il legame lineare esistente tra l'indicatore, che deriva da un modello analogico, e i punteggi fattoriali delle unità, calcolati con il metodo delle componenti principali. Questa relazione consente, da un lato, l'utilizzazione dell'indicatore come un utile strumento complementare nell'analisi delle componenti principali, e dall'altro, di ottenere, proprio da questa metodologia, una interpretazione diretta dell'indicatore. Il criterio è illustrato con l'ausilio del caso concreto di un film per arti grafiche.

Keywords: Multivariate Statistical Analysis, Performance Indices, Total Quality.

1. Introduction

Let be j ($j = 1, 2, \dots$) the j^{th} realization of an n output process and i ($i = 1, 2, \dots, n$) its i^{th} output. Let be $v_i^{(j)}$ a normalized performance measurement, on an interval scale, of the i^{th} output in the j^{th} realization and γ_i ($0 < \gamma_i \leq 1$) an importance weight of the i^{th} output. Let be $x_i^{(j)} = \gamma_i v_i^{(j)}$; the quantities γ_i and $x_i^{(j)}$ may be viewed as the components of two vectors, say γ and $\mathbf{x}^{(j)}$. The moduli of the two vectors may represent, respectively, the measures of a perfect realization and that of a j^{th} generic one. Hence, a first quantity, $c_1^{(j)}$, defined as the ratio between the moduli of $\mathbf{x}^{(j)}$ and γ may represent the overall performance of the j^{th} realization, expressed in unit fractions:

$$c_1^{(j)} = \sqrt{\frac{\sum_{i=1}^n x_i^{(j)2}}{\sum_{i=1}^n \gamma_i^2}}$$

Further, because any angle $\theta^{(j)} \neq 0$, between $\mathbf{x}^{(j)}$ and γ indicates lack of proportionality between the actual individual performances and corresponding importance weight, a second quantity, $c_2^{(j)}$, has been connected to $\theta^{(j)}$. Considering that this angle has a maximum, say, $\theta_{\max}$, and that the following relations hold true, (Cascini, 1995),

$$\theta_{\max} = \arccos \frac{\min(\gamma_1, \dots, \gamma_i, \dots, \gamma_n)}{\sqrt{\sum_{i=1}^n \gamma_i^2}} ; \quad \theta^{(j)} = \arccos \frac{\sum_{i=1}^n x_i^{(j)} \gamma_i}{\sqrt{\sum_{i=1}^n x_i^{(j)2} \sum_{i=1}^n \gamma_i^2}}$$

the second quantity, $c_2^{(j)}$ is defined by the following relation:

$$c_2^{(j)} = 1 - \frac{|\theta^{(j)}|}{|\theta_{\max}|}$$

which may indicate the adequacy of single factor performances to its importance weight. Incidentally, a way to use the indicator, with a practical application, can be found in (Cascini, 1995). A third quantity, say $I_q^{(j)}$, can be defined on the basis of an analogical model; in fact, assuming that a perfect realization of the process is equivalent to the energy $\sum_{i=1}^n \gamma_i^2$ of a *system of n unitary masses, with common speed $\gamma_i \sqrt{2}$* , and that the performance of a generic process realization j , to the energy $\sum_{i=1}^n v_i^{(j)} \gamma_i^2$ of a *system of n masses, of size $v_i^{(j)} \leq 1$, with same speed $\gamma_i \sqrt{2}$* , the ratio between these energies can be assumed to indicate the performance level of the j^{th} process realization, as perceived by the end user, i.e.

$$I_q^{(j)} = \frac{\gamma_1^2}{\sum_{i=1}^n \gamma_i^2} v_1^{(j)} + \dots + \frac{\gamma_i^2}{\sum_{i=1}^n \gamma_i^2} v_i^{(j)} + \dots + \frac{\gamma_n^2}{\sum_{i=1}^n \gamma_i^2} v_n^{(j)}$$

It can be verified that the following relation (1) connects the quantities $I_q^{(j)}$, $c_1^{(j)}$, $c_2^{(j)}$.

$$I_q^{(j)} = c_1^{(j)} \cos(\theta_{\max} (1 - c_2^{(j)})) \quad (1)$$

The idea is using the $I_q^{(j)}$ indicator as a subsidiary tool in the principal component analysis, essentially for ranking the statistical units, and conversely, obtaining from the principal component methodology a refined interpretation of its meaning. The study case, illustrated in the section 2, below, is used to show how to reach these objectives.

2. A study case

One roll, of specified size, represents one result of the manufacturing process of polyester film for graphic arts. The process produces an indefinite numerable sequence (1,2,...,j,...) of rolls. A roll quality is determined on the basis of the numerical values

assigned to 9 quality characteristics, $(1,2,...,i,...,9)$, each one included in the range $(0,1)$. The importance weights, γ_i , of each characteristic are also included in the range $(0,1)$. Quality characteristics are: 1. *Electrostatic discharge rate*; 2. *Antistatic amount in the film*; 3. *Silica amount in the film*; 4. *Surface friction coefficient*; 5. *Ink absorbance*; 6. *Thickness*; 7. *Surface defects*; 8. *Haze*; 9. *Heat shrinkage*. The numerical value assigned to the i^{th} ($i=1,2,...,9$) characteristic, within the j^{th} roll, is reported as $v_i^{(j)}$. In Table 1 are reported the numerical values assigned to the 9 characteristics, within 10 different rolls, as well as the values of the overall performance indicators $c_1^{(j)}, c_2^{(j)}, I_q^{(j)}$, defined above. Considering the quantities $x_i^{(j)} = \gamma_i v_i^{(j)}$, 10 vectors, of size 9, can be defined, say $\mathbf{x}_{(j)} = (x_1^{(j)}, x_2^{(j)}, ..., x_i^{(j)}, ..., x_9^{(j)})'$; ($j=1,2,...,10$), which are 10 statistical units or 10 realizations of the random vector, say $\mathbf{x}_{(9)} = (x_1, x_2, ..., x_i, ..., x_9)'$. In Table 2 are reported, the 9 coefficients of the first 4 principal components, $\beta_x^{(s)}$; ($s=1,2,3,4$), based on covariance, which explain the 91.6% of total variance, and scores of the 10 j units, for the first 4 factors. The score for the factor s of the statistical unit j is indicated as:

$$\zeta_s^{(j)} = \sum_{i=1}^9 \beta_x^{(s)} x_i^{(j)}$$

Table 1: 10 results of the manufacturing process of a polyester roll, for graphic arts. A roll quality is identified by the $v_i^{(j)}$ values given to 9 different quality characteristics.

Process result $v_i^{(j)}$											
i	γ_i	$v_i^{(1)}$	$v_i^{(2)}$	$v_i^{(3)}$	$v_i^{(4)}$	$v_i^{(5)}$	$v_i^{(6)}$	$v_i^{(7)}$	$v_i^{(8)}$	$v_i^{(9)}$	$v_i^{(10)}$
1	0.6	0.7	0.8	1.0	0.4	0.9	0.2	0.5	0.9	0.7	0.4
2	0.5	0.6	0.7	1.0	0.5	0.8	0.3	0.5	0.8	0.6	0.5
3	0.8	0.5	0.8	0.7	0.6	0.7	0.4	0.4	0.6	0.3	0.3
4	1.0	0.9	0.2	0.6	0.1	0.4	0.4	0.2	0.1	0.6	1.0
5	0.8	0.7	0.7	0.9	0.5	1.0	0.2	0.9	0.7	0.8	0.3
6	0.6	1.0	0.2	0.5	0.9	1.0	1.0	0.7	0.8	0.3	0.5
7	1.0	0.4	0.2	0.7	0.9	0.7	0.9	1.0	0.2	0.4	0.7
8	1.0	0.9	0.9	1.0	0.5	0.3	0.2	0.4	0.9	1.0	1.0
9	1.0	1.0	0.6	0.8	0.7	0.7	0.9	0.5	1.0	1.0	0.4
	$c_1^{(j)}$	0.788	0.612	0.802	0.615	0.692	0.612	0.619	0.702	0.725	0.690
	$c_2^{(j)}$	0.794	0.652	0.854	0.688	0.748	0.608	0.672	0.631	0.725	0.691
	$I_q^{(j)}$	0.757	0.544	0.786	0.559	0.651	0.526	0.558	0.615	0.674	0.629

From preceding observations we get, in general, that, for an n output process, $n+1$ linear relations, i.e., $I_q = I_q(x_1, ..., x_i, ..., x_n)$; $\zeta_p = \zeta_p(x_1, ..., x_i, ..., x_n)$, ($p=1,2,...,n$), can be written. Solving the ζ_p system, n linear functions, $x_i = x_i(\zeta_1, ..., \zeta_p, ..., \zeta_n)$, ($i=1,2,...,n$) are obtained. Inserting them into relation $I_q = I_q(x_1, ..., x_i, ..., x_n)$, we can conclude that the relation $I_q = I_q(\zeta_1, ..., \zeta_p, ..., \zeta_n)$ is a linear function. In the case under study we have obtained the following relation (2).

$$I_q = 0.121\zeta_1 - 0.049\zeta_2 - 0.046\zeta_3 + 0.279\zeta_4 + 0.078\zeta_5 - 0.221\zeta_6 - 0.036\zeta_7 - 0.078\zeta_8 + 0.029\zeta_9 \quad (2)$$

The I_q indicator can be used to rank the statistical units, because $c_1^{(i)}, c_2^{(i)}$ are linked to process performance, and I_q is linked to $c_1^{(i)}, c_2^{(i)}$, by means of relation (1). Conversely, few latent significant variables, (4 in our case), can facilitate interpreting the I_q sense.

Table 2: Coefficients of the first 4 principal components of $\mathbf{x}_{(9)}$ and scores for the first principal components of the 10 statistical units (polyester rolls).

i	x_i	$\beta_x^{(1)}$	$\beta_x^{(2)}$	$\beta_x^{(3)}$	$\beta_x^{(4)}$	j	$\zeta_1^{(j)}$	$\zeta_2^{(j)}$	$\zeta_3^{(j)}$	$\zeta_4^{(j)}$
1	x_1	0.2161	0.2582	0.0782	0.3547	1	0.8783	-0.300	-0.605	1.2489
2	x_2	0.1219	0.1408	0.0959	0.2251	2	0.8483	0.3283	-0.031	0.7728
3	x_3	0.0313	0.2848	0.1032	0.1085	3	0.8191	-0.032	-0.093	1.3081
4	x_4	0.3189	-0.761	-0.114	0.4262	4	-0.003	0.0596	-0.243	0.8893
5	x_5	0.1096	0.3346	0.2476	0.6581	5	0.1791	0.1706	-0.253	1.4277
6	x_6	-0.237	-0.018	-0.436	0.2743	6	-0.168	-0.275	-0.610	0.8726
7	x_7	-0.537	-0.317	0.2832	0.2706	7	-0.049	0.0081	0.0283	1.1333
8	x_8	0.6785	-0.108	0.1928	-0.180	8	0.7997	0.4421	-0.491	0.8946
9	x_9	0.1365	0.1682	-0.769	0.1264	9	0.9539	-0.094	-0.365	1.0231
						10	0.7211	-0.784	-0.034	0.8935

3. Concluding Remarks

Ranking multivariate sample units of an n output process can be obtained by means of values assumed by a simple indicator of their overall perceived performances. The indicator can be used as a subsidiary tool in the analysis of principal components. Conversely, the results of the principal component analysis, which, usually, consist in a reduction of the explicative variables of a process, can simplify its interpretation, making clear the relevant reasons for its variability.

References

- Cascini E. (2001) Un indicatore della qualità complessiva percepita, *Qualità*, 9, 42-46
- Cascini E. (1995) Produzione e Tecnologia: una testimonianza aziendale, *Il Quaderno dell'Istituto Guglielmo Tagliacarne*, 9, 101-120
- Anderson T.W. (1958) *An Introduction to Multivariate Statistical Analysis*, Wiley, New York
- Kendall M., Stuart A., Ord J.K. (1983) *The Advanced Theory of Statistics*, Volume 3, Charles Griffin Co Ltd
- Fabbris L. (1997) *statistica multivariata, analisi esplorativa dei dati*, McGraw-Hill Libri Italia srl