Composite Index

Rationale & Methodology

The script allows estimating a Composite Index (C) that maximizes the Fisher's distance between 2 distributions of 2 sub-sample of C selected according to a conditional variable.

- The composite index is defined as the algebraic sum of K variables:

$$C = \sum_{k=1}^{K} X_k \qquad (1)$$

  where X is the array of data for a given variable expressed as Z-score (i.e., mean=0 and standard deviation =1). Note that because the variables are standardized as Z-score variables addition in C is allowed.

- The Distance (D) between two subsample of C is computed using Fisher's distance D (Fisher 1936):

$$D = \frac{(\mu_1 - \mu_2)^2}{(\sigma_1 - \sigma_2)^2} \qquad (2)$$

  where μ_1 and μ_2 are the distribution mean values of the two sub-sample of C, σ_1 and σ_2 their corresponding standard deviations.

- The conditional variable is used to define a Boolean variable to select the two-subsample of C. Suppose that 10 variables are used, each one with 100 observations. Observations must be coherent between variables, in other words, the first observation of each variable must refer to the same area/year/species. Similarly, the same must hold for the other remaining 99 observations. The conditional variable will say which observation (i.e., rows) belong to the first and second groups, respectively.

The rationale behind the composite index approach is that if the distance computed for a composite index (^{C}D) given by the sum of two variables, is larger than the distance computed for each variable separately then we can infer that the joint effect of the two components is more effective than the individual components in segregating the 2 distributions defined by the conditional variable. However, interactions between components may reduce, increase or have no effect on ^{C}D, hence the goal is to include only those variable Z-scores that maximize ^{C}D.

The recursive calculation identifies the variables that combined additively as a composite index (C, eqn. 1), maximize the value of D (i.e., maximize the objective function of eq. (2)).

Specifically, to determine C that maximizes ^{C}D, we apply the following recursive procedure:

i) set C equal to a single variable and evaluated the related distance ^{C}D between its distributions under exceptionally high and low yields according to Eqn. (2).

ii) use a second variable from the remaining K-1 ones that maximizes ^{C}D when added to the C calculated at step (*i*). Note that the addition of a variable is tested both by summing and subtracting the Z-scores to determine the stress direction of the variables. Note that in some cases, variables need to be subtracted to maximize ^{C}D rather than added. In such cases the variables are indicated with a minus sign (-).

iii) repeat step *(ii)* K-2 times until all the independent variables were included in C to identify the sequence of variables that underlie a monotonic increase of ^{C}D for C.

iv) because the results may depend on the starting variable, procedures (*i-iii*) are repeated K times starting each time with a new initial variable from the list of K variables.

At the end of the recursive procedure, K curves may be drawn for the values of ^{C}D consisting of the added components in C, where each curve starts with a different initial variable. The one yielding the largest monotonic increase of ^{C}D is identified, but more than one could result in tie cases. The underlying component variables of the monotonic increase in ^{C}D define the composite risk index.

Data used in the example

Data used in the script uses 23 climate-related variables (i.e. K=23, see eqn. 1) as inputs and Italian olive yields as the conditional variable. Climate-related variables refer to the period 2006-2020, aggregated at the provincial level according to the metric reported in the last column of Table 1. Yield data were retrieved from the Italian National Statistics Institute (ISTAT) whereas climate data from the ERA5 dataset, available at the Copernicus Climate Database of the European Center for Medium-Range Weather Forecast (ECMWF).

Yields and climate-related variables are organized into synthetic matrices (*X* in equation 1) of N provinces with a complete time-series by M years (i.e., N=66, M=15). A large number of variables are aggregated on a bimonthly time scale. The six bimonthly periods are January-February, March-April, May-June, July-August, September-October, and November-December, denoted by subscripts "*_bx*" where x = 1,…,6 is the ordinal integer of the bimonthly interval. One variable is aggregated on an annual basis. Variables computed for bimonthly periods are indicated by "b" in the last column of Table 1, while those aggregated on an annual basis are indicated by "a". For example, the variables expressing the annual count of days when daily precipitation is larger than 10mm (RR10) and the minimum temperature (Tmin) of the second bimonthly period (b2) are *RR10_a* and *Tmin_b2*, respectively. Original climate-related variables were much more than 23 but have been reduced by feature selection by the author. Metadata on climate-related variables are shown in Table 1.

The file main.py will run the Cindex function (coded into the library auxiliary.py) using the data on climate-related variables and olive yields provided along with the scripts. To proper install and run the script read the instructions on the README.MD file.

Table 1. List of variables.

acronym	Long name	unit	time scale and metric
Tmin	Minimum air temperature at 2 meters	°C	b, minimum
Tmax	Maximum air temperature at 2meters	°C	b, maximum
Tave	Temperature at 2 meter	°C	b, average
SSR	Surface net solar radiation	J m^{-2}	b, average of daily maximum
RH	Relative humidity	0-100	b, average
DTR	Daily temperature range	°C	b, average
R5max	maximum 5-consecutive-days precipitation	dimensionless	b, maximum
CDD	Maximum length of dry spell	dimensionless	b, maximum
GDD	growing degree day above 0°C	°C	b, cumulated
RR10	annual count of days when daily precipitation is larger than 10mm	dimensionless	a