Construction of composite indices in presence of outliers

Mishra, SK

North-Eastern Hill University, Shillong (India)

26 May 2008

Construction of Composite Indices in Presence of Outliers

SK Mishra
Dept. of Economics
North-Eastern Hill University
Shillong (India)

I. Introduction: Oftentimes we require constructing composite indices by a linear combination of a number of indicator variables. If we denote the indicator variables by $X=[x_1, x_2, ..., x_m]$ where each x_j has n observations (cases) and weights assigned to those variables by $w=[w_1, w_2, ..., w_m]'$ then the composite index $I = Xw$ obtains a single value for each case k, or $I_k = \sum_{j=1}^{m} x_{kj} w_j;\ k=1,n$. The weights may be determined subjectively or objectively by certain considerations extraneous to the dataset X, or alternatively they may endogenously be determined by the statistical information obtained from dataset X itself. Endogenous weights are frequently obtained by a statistical technique called the Principal Components Analysis (PCA), which maximizes the sum of squared coefficients of (the product moment) correlation between the derived composite index I and the indicator variables, X, or stated differently, $I = Xw$ such that $\sum_{j=1}^{m} r^2(I, x_j)$ is maximum.

In presence of sizeable outliers in the data variables, X, we cannot expect the product moments correlation coefficients to remain unaffected. The outliers distort mean, standard deviation and the covariance structure of the indicator variables leading to distortion in the coefficient of correlation. It may be desirable, therefore, to devise a technique that would minimize the influence of outliers on the composite index. Our objective in this paper is to propose a new technique to construct such a composite index. We also demonstrate the effectiveness of the proposed technique by a simulation experiment.

II. The Coefficient of Correlation in the Median Family: It is well known that median as a measure of central tendency is (normally) unaffected by the presence of outliers in the data. The median is an analogue of the (arithmetic) mean; it minimizes the sum of probability-weighted absolute deviations of data points from itself ($\min_c \left| \sum_{i=1}^{n} |x_i - c|^L p_i \right|^{1/L}$ for L=1) while the arithmetic mean minimizes the probability-weighted sum of squared deviations of data points from itself (that implies $\min_c \left| \sum_{i=1}^{n} |x_i - c|^L p_i \right|^{1/L}$ for L=2).

Bradley (1985) showed that if $(u_i, v_i);\ i=1,n$ are n pairs of values such that the variables u and v have the same median = 0 and the same mean deviation (from median) or $(1/n)\sum_{i=1}^{n}|u_i| = (1/n)\sum_{i=1}^{n}|v_i| = d \neq 0$, both of which conditions may be met by any pair

of variables when suitably transformed, then the absolute correlation may be defined as $\rho(u,v) = \sum_{i=1}^{n}\left(|u_i+v_i|-|u_i-v_i|\right)\Big/\sum_{i=1}^{n}\left(|u_i|+|v_i|\right)$.

III. Construction of a Composite Index Using Bradley's Correlation: Bradley's coefficient of correlation (that belongs to the median family) is an analogue of the Pearson's product moment correlation coefficient (in the family of arithmetic mean). It appears therefore that one may construct a composite index my maximization of the sum of absolute values of Bradley's coefficient of correlation between the composite index, I and the indicator variables. This is to say that we can obtain $I_1 = Xw_1$ such that $\sum_{j=1}^{m}\left|\rho(I_1,x_j)\right|$ is maximal. This composite index, I_1, will be analogous to the PCA-based index, I_2, that maximizes the sum of squared sum of the Pearson's coefficients of correlation between the composite index and the indicator variables or $I_2 = Xw_2 : \max\sum_{j=1}^{m} r^2(I_2,x_j) \Rightarrow \max\left[\sum_{j=1}^{m} r^2(I_2,x_j)\right]^{1/2}$.

IV. Issues Relating to Maximization: Obtaining the PCA-based composite index is simpler since it has a closed form formula. The (Pearson's) correlation matrix, R is constructed from X such that $R = (1/n)X'X$ where $x_j \in X \ \forall\ j$ has zero mean and unit standard deviation. The largest eigenvalue (λ) and the associated eigenvector (e) of R is obtained. The eigenvector is normalized so that $\|e\| = 1$. The normalized eigenvector is used as the weight, w_2, to obtain $I_2 = Xw_2$. It is possible, nevertheless, to directly obtain the composite index, I_2, by maximizing $\sum_{j=1}^{m} r^2(I_2,x_j) : I_2 = Xw_2$.

There is no closed form formula for obtaining $I_1 = Xw_1$ such that $\sum_{j=1}^{m}\left|\rho(I_1,x_j)\right|$ is maximal. Hence, one has to directly obtain it by solving the intricate maximization problem.

V. Nonlinear Optimization by Differential Evolution: The method of Differential Evolution (DE) is one of the most powerful self-organizing, evolutionary, population-based and stochastic global optimization methods. It is an outgrowth of the Genetic Algorithms. The crucial idea behind DE is a scheme for generating trial parameter vectors. Initially, a population of points (p in d-dimensional space) is generated and evaluated (i.e. f(p) is obtained) for their fitness. Then for each point (p_i) three different points (p_a, p_b and p_c) are randomly chosen from the population. A new point (p_z) is constructed from those three points by adding the weighted difference between two points ($w(p_b-p_c)$) to the third point (p_a). Then this new point (p_z) is subjected to a crossover with the current point (p_i) with a probability of crossover (c_r), yielding a candidate point, say p_u. This point, p_u, is evaluated and if found better than p_i then it replaces p_i else p_i remains. Thus we obtain a new vector in which all points are either better than or as good as the current points. This new vector is used for the next iteration. This process makes the differential evaluation scheme completely self-organizing. This method has been successfully applied for optimizing extremely nonlinear and multimodal functions (Mishra, 2007a, 2007b and 2007c).

VI. A Simulation Experiment: We have conducted a simulation experiment to examine the effectiveness of our proposed method. We have generated a matrix, X, of six variables, each in 30 observations. The correlation matrix of these variables is given in Table-1. Using these variables, we have obtained two composite indices by direct optimization: the one (I_{10}) relating to the method proposed by us and the other (I_{20}) relating to the PCA. Both of these indices are standardized by using the relationship $[I_k - \min_k(I_k)]/[\max_k(I_k) - \min_k(I_k)]$; $k = 1,\ n$ so as to make the index values lie between zero and unity These composite indices serve as reference since X does not contain outliers.

It is interesting to note (see table-1) that I_{10} and I_{20} are highly correlated (r = 0.99759), although Bradley weights (w_1) and correlation coefficients (ρ) are uniformly smaller (in magnitude) than the Pearson weights (w_2) and correlation coefficients (r).

Next, we introduce outliers to X. Three outliers (ranging between -10 to 10) have been added to each indicator variable ($x_{j;\ j=1,\ n}$) at random locations. Then, using these (contaminated) variables, the two composite indices (I_{11} and I_{21}) have been obtained. The indices have been standardized as before to lie between zero and unity. The results are presented in Table-2. All derived composite indices are presented in Table-3.

The root-mean-square (RMS) $= \sqrt{(1/n)\sum_{k=1}^{n}(I_{k10} - I_{k11})^2} = 0.062108$ for our proposed method vis-à-vis RMS $= \sqrt{(1/n)\sum_{k=1}^{n}(I_{k20} - I_{k21})^2} = 0.077333$ obtained for the PCA-based index suggests us that in presence of outliers our proposed method will perform better. As shown in the graph (Fig.1), the fluctuations in I_{21} appear to be more than those in I_{11}.

Table.1 : Correlation Coefficients and Weights for the Reference Indicator Variables (Without Outliers)

Variables	X_1	X_2	X_3	X_4	X_5	X_6	I_{10}	I_{20}
X_1	1.00000	0.91112	0.79774	-0.80408	0.90597	-0.88239	0.98312	0.97480
X_2	0.91112	1.00000	0.61258	-0.70371	0.89051	-0.76986	0.91918	0.89877
X_3	0.79774	0.61258	1.00000	-0.76991	0.66145	-0.77614	0.82477	0.84912
X_4	-0.80408	-0.70371	-0.76991	1.00000	-0.82274	0.69284	-0.86607	-0.88139
X_5	0.90597	0.89051	0.66145	-0.82274	1.00000	-0.78670	0.94423	0.93179
X_6	-0.88239	-0.76986	-0.77614	0.69284	-0.78670	1.00000	-0.88785	-0.90247
I_{10}	0.98312	0.91918	0.82477	-0.86607	0.94423	-0.88785	1.00000	0.99759
I_{20}	0.97480	0.89877	0.84912	-0.88139	0.93179	-0.90247	0.99759	1.00000
Bradley weights	0.45546	0.31762	0.32684	-0.29143	0.35443	-0.16293	I_{10} = Composite Index by maximization of the sum of absolute Bradley's Correlation Coefficients	
Bradley Correlation	0.89741	0.75791	0.70183	-0.68475	0.78322	-0.75640		
Pearson weights	0.52399	0.58743	0.78502	-0.85376	0.56337	-0.60604	I_{20} = Composite Index by maximization of the sum of squared Pearson's Correlation Coefficients	
Pearson correlation	0.97480	0.89876	0.84912	-0.88139	0.93179	-0.90247		

Table.2 : Correlation Coefficients and Weights for the Reference Indicator Variables (With three Outliers between -10 and 10)								
Variables	X_1	X_2	X_3	X_4	X_5	X_6	I_{11}	I_{21}
X_1	1.00000	0.68901	0.63464	-0.60439	0.86492	-0.74930	0.96985	0.95590
X_2	0.68901	1.00000	0.53335	-0.23724	0.63100	-0.45318	0.73477	0.74577
X_3	0.63464	0.53335	1.00000	-0.28127	0.48497	-0.45498	0.65326	0.71325
X_4	-0.60439	-0.23724	-0.28127	1.00000	-0.60731	0.45490	-0.57757	-0.67029
X_5	0.86492	0.63100	0.48497	-0.60731	1.00000	-0.60940	0.94002	0.88337
X_6	-0.74930	-0.45318	-0.45498	0.45490	-0.60940	1.00000	-0.76137	-0.78323
I_{11}	0.96985	0.73477	0.65326	-0.57757	0.94002	-0.76137	1.00000	0.97581
I_{21}	0.95590	0.74577	0.71325	-0.67029	0.88337	-0.78323	0.97581	1.00000
Bradley weights	0.35778	0.09415	0.13863	0.04825	0.51405	-0.15286	I_{11} = Composite Index by maximization of the sum of absolute Bradley's Correlation Coefficients	
Bradley Correlation	0.87477	0.65153	0.56840	-0.50193	0.80043	-0.68208		
Pearson weights	0.44192	0.53316	0.68256	-0.66708	0.54588	-0.53644	I_{21} = Composite Index by maximization of the sum of squared Pearson's Correlation Coefficients	
Pearson correlation	0.95590	0.74577	0.71325	-0.67029	0.88337	-0.78323		

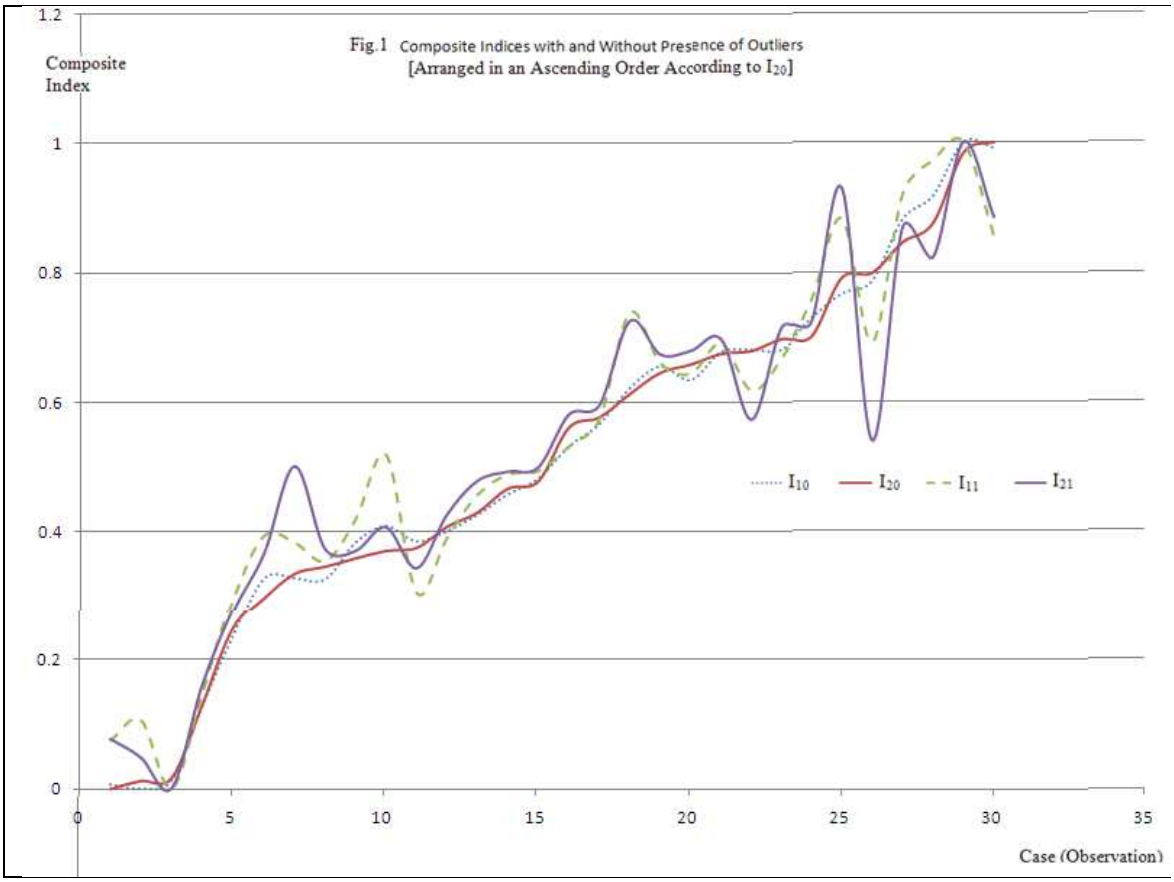

Table.3 : Composite Indices with (-10, 10 range) Outliers and Without Outliers									
	Without Outliers		With Outliers			Without Outliers		With Outliers	
Sl. No.	I_{10}	I_{20}	I_{11}	I_{21}	Sl. No.	I_{10}	I_{20}	I_{11}	I_{21}
1	0.00000	0.01236	0.10662	0.04767	16	0.01245	0.01645	0.00000	0.00000
2	0.23418	0.24737	0.29661	0.27640	17	0.53109	0.56071	0.53143	0.58084
3	0.88073	0.84500	0.92008	0.86838	18	0.63358	0.65802	0.64426	0.67895
4	0.68067	0.67890	0.61788	0.57269	19	0.72741	0.69887	0.75561	0.72107
5	0.76524	0.78990	0.88226	0.92897	20	0.65483	0.64521	0.66060	0.67342
6	0.38436	0.37486	0.30520	0.34277	21	0.32729	0.33457	0.38292	0.50055
7	0.00632	0.00000	0.07506	0.07659	22	0.62112	0.61233	0.73851	0.72505
8	0.32555	0.34553	0.35433	0.37221	23	0.45723	0.46635	0.48820	0.49213
9	0.12642	0.12814	0.14552	0.15830	24	0.32696	0.29637	0.39360	0.36586
10	0.48163	0.47763	0.49373	0.49912	25	0.78514	0.79752	0.69088	0.53828
11	0.68082	0.69761	0.66917	0.71659	26	0.42541	0.42938	0.45679	0.47811
12	0.38275	0.35851	0.41909	0.36965	27	0.40770	0.37009	0.51886	0.40637
13	0.56575	0.57621	0.57338	0.59467	28	0.91677	0.87460	0.97238	0.82189
14	0.40016	0.40747	0.39010	0.42586	29	0.99074	1.00000	0.85489	0.88409
15	1.00000	0.98364	1.00000	1.00000	30	0.67744	0.67450	0.69074	0.69654

References

Bradley, C. (1985) “The Absolute Correlation”, *The Mathematical Gazette*, 69 (447), pp. 12-17.

Mishra, S.K. (2007a): “Performance of Differential Evolution Method in Least Squares Fitting of Some Typical Nonlinear Curves” *Journal of Quantitative Economics*, Vol. 5 (1), pp.140-177.

Mishra, S.K. (2007b): “Least Squares Estimation of Joint Production Functions by the Differential Evolution method of Global Optimization.” *Economics Bulletin*, Vol.3 (51), pp.1-13

Mishra, S.K. (2007c) "Construction of an Index by Maximization of the Sum of its Absolute Correlation Coefficients with the Constituent Variables" SSRN: http://ssrn.com/abstract=989088

Note: A Fortran Computer program to compute Composite Indices using Bradley’s absolute correlation and PCA by direct maximization is available on http://www.webng.com/economics