Hindawi
Journal of Probability and Statistics
Volume 2022, Article ID 4155384, 12 pages
https://doi.org/10.1155/2022/4155384

Research Article

Interpretability of Composite Indicators Based on Principal Components

Kris Boudt,[1,2,3] **Marco d'Errico**,[4,5] **Hong Anh Luu**,[2,4] **and Rebecca Pietrelli**[4]

[1]*Department of Economics, Ghent University, Ghent, Belgium*
[2]*Department of Business, Vrije Universiteit Brussel, Brussel, Belgium*
[3]*School of Business and Economics, Vrije Universiteit Amsterdam, Amsterdam, Netherlands*
[4]*Food and Agriculture Organization of the United Nations, Rome, Italy*
[5]*La Sapienza University of Rome, Rome, Italy*

Correspondence should be addressed to Kris Boudt; kris.boudt@ugent.be and Hong Anh Luu; hong.anh.luu@vub.be

Received 9 May 2022; Revised 11 August 2022; Accepted 24 August 2022; Published 29 September 2022

Academic Editor: Muhammad Ahsan

Principal component approaches are often used in the construction of composite indicators to summarize the information of input variables. The gain of dimension reduction comes at the cost of difficulties in interpretation, inaccurate targeting, and possible conflicts with the theoretical framework when the signs in the loading are not aligned with the expected direction of impact. In this study, we propose an adjustment in the construction of principal component approaches to avoid these problems. The effectiveness of the proposed approach is illustrated in defining the Food and Agriculture Organization of the United Nations' Resilience Capacity Index, which is used to measure household-level resilience to food insecurity. We conclude that the robustness gain of using the new method improves the reliability of the composite indicator.

1. Introduction

Composite indicators are widely used to evaluate the overall performance of entities on multiple criteria [1]. Their rising popularity comes from the power of simplification by summarizing a complex and sometimes elusive process into a single figure. Constructing a composite indicator requires carefully analyzing the underlying structure of the input variables, as it helps assess the suitability of the data and provides an understanding of the implications of the methodological choices [2]. Various techniques can be used in this step, among which principal component analysis is a popular choice. Its related factor analysis method is most preferred in the development of composite indicators, as this approach is simple and allows for the construction of weights representing the information content of individual indicators [2] (In this paper, we refer to "principal component analysis" and "principal component factor analysis" generally as "principal component approaches"). There are multiple aggregate measures that employ principal component approaches in their constructions. Some examples are the product market regulation index of [3], the sustainability indicator proposed by [4], and the wealth index of [5], which is used in Demographics and Health Surveys (DHS) reports [6] and UNICEF Multiple Indicator Cluster Surveys [7].

Composite indicators constructed using traditional principal component approaches can encounter some potential limitations. Without loss of generality, we assume to have a set of input variables in which they are all defined in such a way that an increase leads to an improvement of the state proxied by the aggregate index. Under this assumption, we expect the composite index to positively correlate with all of its sub-indicators. However, negative signs in the classical loading matrix can reverse this relation, dampening the representative power of the aggregate indicator. They can also deflect how changes in input variables should be loaded to the composite indicator, leading to inaccurate ranking and targeting of entities' performances. Moreover, the traditional loading solution includes both positive and negative values, which can result in interpretation difficulty. If two indicators have the same sign on one loading vector, they tend to increase or decrease together on

that score. If they have opposite signs, there is a trade-off between these two variables [8]. When an individual variable has sizable positive loading in one score and significant negative loading in another [2], it can be confusing and difficult to conclude the overall direction of impact.

This study proposes a solution for the above problems by modifying the traditional principal component approaches to consider the expected direction of impact while grouping common information from the data set. Using an alternative version of the loading matrix that contains only positive (or negative) elements, we control how variables influence principal component scores, thereby ensuring accurate targeting and ranking of entities' performance and a consistent relationship between input indicators, scores, and the composite indicator. Additionally, the new loading solution reallocates the concentration of elements to provide a straightforward interpretation. Overall, our method supports the reading of results and the accuracy of the estimated indicator.

The methodological approach taken in this study is based on optimization techniques. As explained by [8], outputs of classical principal component analysis (PCA) correspond to optimal solutions of several optimization problems simultaneously, and we can consider possibilities for tweaking these allocation problems to find "better" solutions that satisfy alternate criteria. Following this inspiration, we first develop an optimization-based process to extract loading vectors of the traditional PCA. Then, to acquire a new matrix with only positive (or negative) values, zero lower (or upper) bounds are enforced to the original allocation problems. The obtained solution is also applicable in a factor model framework. Principal component factor analysis (PCFA) differs from PCA in terms of score estimation, as the model assumes that the data are based on the underlying latent factors. The communalities in this factor model are assumed to be 1. Thus, the estimators of the latent variables obtained are proportional to those given by the PCA model [9]. Using unrotated PCA loadings as pattern coefficients [10], we build up formulas for classical and restrained PCFA approaches.

To date, much research focused on creating new methods for PCA and factor analysis (FA) based on the particular analysis needed. Common goals are to increase the robustness to outlier observations of PCA [11, 12] and FA [13–15]; to create sparse principal components where the loadings have few nonzero entries [16, 17]; or to create smooth scores where the loadings of certain variables are similar [18]. There are also several studies investigating the nonnegativity of PCA and FA loading solutions with different techniques such as oblique factor rotation with partially nonnegative constraint [19], positive matrix factorization [20, 21], or singular value decomposition with modifications in exponential parameterization [22]. This research contributes to the existing literature using optimization techniques to find the loading that maximizes the variance explained under the constraint that the signs of the loadings are as expected.

Although from the methodology we have solutions for both restrained PCA and PCFA, the application section only demonstrates and discusses in detail the impact of constraints under PCFA. There are two main reasons for this setting. First, PCA and PCFA return similar or proportional results. Key conclusions about the impact of constraints obtained from the two methods are the same; thus, it is a repetition to explain them both. Second, the resilience pillars employed for demonstration in this section are normally constructed using FA [23]; therefore, discussing PCFA results is more relevant for practical usage.

To compare the performance of the new approach over the classical one, we use them in the construction of the four pillars of the Resilience Capacity Index (RCI), used by the Food and Agriculture Organization (FAO) to understand how households cope with shocks and stressors. This is a measure of performance that assesses the resilience of households and is built on four fundamental attributes—namely, access to basic services, assets, social safety net, and adaptive capacity. Each of these pillars itself is a composite indicator aggregated from a subset of selected indicators.

The theoretical framework suggests that a pillar should have positive relationships with all of its input variables. Using two data sets collected from Uganda in 2019 and Chad in 2014, we find that this relationship is not always guaranteed under classical PCFA. The mixed-sign loading matrices can result in pillars that have negative correlations with one of their sub-variables. Moreover, some variables have significant positive loading in one component and significant negative loading in another, making it difficult to conclude the overall influence on the composite indicator. By design, the constrained solution does not suffer from this lack of interpretability. The constraints ensure that all loading elements are nonnegative. By construction, we thus have a positive correlation between a pillar and each input variable. As the signs in the loading are consistent and aligned with the expected direction, the changes in input variables are accurately reflected in values of the pillar and therefore lead to better targeting and changes in ranks of household performance.

The remainder of the study is organized as follows. Section 2 describes our motivation. Section 3 presents the methodology, while Section 4 discusses the application results. Section 5 contains our conclusions.

2. Motivating Example

In what follows, we use a simplified numerical example to illustrate the problems under traditional principal component approaches and how our proposed constraints fix them. The illustration is made based on the theoretical framework of *access to basic service* pillar, which is fully explained in Section 4.

Assume that after a project deployed to support a local community, we interview a number of households and obtain data for four input variables: *access to clean water, improved sanitation, closeness school,* and *closeness hospital*. Variables *access to clean water* and *improved sanitation* show whether a household can use water from clean sources and improved toilet facilities, while *closeness school* and *closeness hospital* indicate how close (in physical distance) the family

is located relative to the nearest school and hospital. All of them are structured so that higher values indicate better performances, and we want to extract their underlying overall index *access to basic service*. The theoretical framework suggests that an increase in *access to clean water, improved sanitation, closeness school,* and *closeness hospital* should lead to an improvement in household accessibility to basic service. Thus, we expect the composite indicator *access to basic service* to have positive correlations with all of its input indicators.

In Table 1, we use the traditional and constrained version of PCFA to analyze the input data. Following the Kaiser criterion, we drop all factors with eigenvalues below 1.0 [2, 24] and keep the first two scores to construct the overall index. The left panel of Table 1 (entitled "Original loadings") presents the loadings of the scores obtained using the traditional PCFA approach (without constraints), together with the correlations between the composite indicator built from them and input variables. We see that the traditional method focuses only on retaining maximum variance and returns a mixed-sign loading matrix. Thus, the composite indicator has a negative correlation with the first sub-indicator, indicating that an increase in the ability to get *clean water* leads to a decrease in *access to basic service*. This conflicts with what is suggested by the above theoretical framework. In contrast, if we use the proposed constrained approach, we obtain the loading of the scores presented in the right panel. Due to the constraints, the loading is kept nonnegative for all elements, ensuring a consistent positive correlation between input variables and the aggregate. Moreover, under the new loading pattern, each sub-indicator is significantly loaded on only one score, giving a clear idea of how it influences the overall index.

3. Methodology

3.1. The Framework. Assume that we have data for n households on m input variables, and they are standardized and stacked into an $n \times m$ matrix named X. These input variables are correlated and are designed such that higher values indicate better performance. Our goal is to obtain an $n \times 1$ vector holding the values of the single composite indicator CI that can represent X.

To get CI, we first apply PCA/PCFA to summarize the underlying structure of X using a new set of q uncorrelated intermediate predictors or scores ($q \leq m$). Then, each of these scores is assigned a weight equal to the proportion of the explained variance [2]. From there, we derive CI by linearly aggregating the scores and their corresponding weights. Let $Z = (z_1, \ldots, z_q)'$ be the $n \times q$ matrix of scores, and $w = (w_1, \ldots, w_q)'$ be the $q \times 1$ weight vector with w_j the share of variance explained by z_j. We then have the general definition of CI as follows:

$$\mathrm{CI} = \sum_{j=1}^{q} z_j w_j = Z^T w. \tag{1}$$

In what follows, we focus on building up the specific expressions for Z and w under different model choices. First, we present how they can be obtained when finding the loadings under traditional principal component approaches. Then, we explain how the new constrained solutions can be derived by adjusting the above procedures.

3.2. Classical Solutions Using Optimization

3.2.1. Under PCA. Under the PCA method, Z is obtained using

$$Z = XV, \tag{2}$$

where X is the original data set, $V = (v_1, \ldots, v_q)'$ the $m \times q$ loading matrix. This process also provides an estimation of $w_j = (1/m)\mathrm{var}(Xv_j)$ [25].

To determine V, we follow [8, 25] using optimization-based techniques. The traditional PCA framework projects the loading vectors such that the first score accounts for the largest possible proportion of the explained variance, the second score explains for the second highest variance share, and so on. This can be done by seeking a vector v_1 such that the variance of projections in X onto this subspace is maximized, and once we have found it, we can then seek a second vector v_2 orthogonal to v_1, which maximizes the variance "left over" after the first projection. The variance of points projected on vector is $\mathrm{var}(Xv_1) = v_1^T S v_1$, where S is the sample covariance/correlation matrix given by $S = (1/n-1)X^T X$ [8]. As such, the first optimization problem can be written as follows:

$$\begin{aligned} &\max_{v_1} v_1^T X^T X v_1, \\ &\text{s.t} \quad v_1^T v_1 = 1. \end{aligned} \tag{3}$$

After finding v_1, we can get the second direction by solving the following problem:

$$\begin{aligned} &\max_{v_2} v_2^T X^T X v_2, \\ &\text{s.t} \quad \begin{cases} v_1^T v_2 = 0, \\ v_2^T v_2 = 1. \end{cases} \end{aligned} \tag{4}$$

Note that the first constraint is linear because we assume that v_1 is known. Here, the condition v_2 to be orthogonal to v_1 is added to enforce the idea that v_2 captures the variation left over after projection on v_1. Carrying on, we can find the k^{th} loading vector by solving

$$\begin{aligned} &\max_{v_k} v_k^T X^T X v_k, \\ &\text{s.t} \quad \begin{cases} v_j^T v_k = 0 \, (j = 1, \ldots, k-1), \\ v_k^T v_k = 1, \end{cases} \end{aligned} \tag{5}$$

where v_k is restricted to be orthogonal to all preceding vectors so that it explains the remaining variance in X after all previous projections. The variance of points projected on vector v_k is $\mathrm{var}(Xv_k) = v_k^T S v_k$.

Table 1: Numeric example to illustrate the explainability of composite indicators.

Input variable ID	Original loadings			Loadings with constraints		
	Loading score 1	Loading score 2	Correlation with CI	Loading score 1	Loading score 2	Correlation with CI
1	−0.54	0.24	−0.33	0.00	1.00	0.62
2	0.53	0.23	0.63	0.67	0.00	0.50
3	0.10	0.92	0.70	0.36	0.00	0.32
4	0.50	−0.17	0.34	0.58	0.00	0.43
Variance explained	1.19	1.01		1.09	1.00	

The vectors v_j with $j = 1, \ldots, q$ form the first q columns of the loading matrix V, while $w_j = (1/m)\mathrm{var}(Xv_j)$ makes the corresponding weight vector w. Applying V to equation (2), we have the estimation of Z under the traditional PCA approach.

3.2.2. From PCA to PCFA. FA is similar to PCA. However, although PCA is based simply on linear data combinations, FA assumes that the data are based on the underlying latent factors [2]. There are a number of different methods to estimate the factor score. Here, we present the formula under PCFA—a special case of the general factor model where the communalities are assumed to be 1. Following [9, 10], estimators of the latent variables obtained in PCFA are proportional to those given by PCA. We can use the results obtained in Section 3.2.1 to estimate the value of Z as follows:

$$Z = D^{-(1/2)}VX, \tag{6}$$

where V is the PCA loading matrix, X is the original data set, and $D = \mathrm{diag}(\mathrm{var}(Xv_1), \ldots, \mathrm{var}(Xv_q))$ is the diagonal matrix of the variance explained. We refer to $D^{-(1/2)}V$ as the loading matrix under this method. The weight vector w can be estimated using the same formula given in Section 3.2.1.

3.3. New Solutions under Constraints. In this subsection, we look for new solutions of Z and w that can fix the problems mentioned in Section 2. Given that all indicators X follow the same structure (Subsection 3.1) and w and $D^{-(1/2)}$ are always positive, the loading matrix V determines the relationships between X, Z, and CI under both PCA and PCFA. Thus, our target narrows down to finding a constrained version of V in which all elements have the same sign. Once the adjusted V is found, we can apply the formula in Subsection 3.2 to obtain the restrained solutions of Z and w.

3.3.1. Methodology. To obtain a same-sign loading matrix V, we add a bound to the optimization function (5) when finding each vector v_k. Although this constraint can be either positive or negative depending on the characteristics of CI, here the positive case is presented. Furthermore, we replace the constraint $v_j^T v_k = 0\,(j = 1, \ldots, k-1)$ in equation (5) with $v_j^T v_k \le \kappa_j$. The threshold value $\kappa_j \ge 0$ can be set to strike a balance between orthogonality of the factors ($\kappa_j = 0$) and flexibility (larger values of κ_j). Thereby, the new solution for the k^{th} loading vector can be obtained by solving

$$\begin{aligned} &\max_{v_k} v_k^T X^T X v_k, \\ &\text{s.t} \begin{cases} v_j^T v_k \le \kappa_j\,(j = 1, \ldots, k-1), \\ v_k^T v_k = 1, \\ v_k \ge 0_m. \end{cases} \end{aligned} \tag{7}$$

Note that the constraint $v_j^T v_k \le \kappa_j\,(j = 1, \ldots, k-1)$ is linear because we assume that all $v_j\,(j < k)$ are known when finding v_k. In this study, we set $\kappa_j = 0$. It follows, then, that whenever a previous factor has a positive loading on a variable, a subsequent factor must have a zero loading. This has the advantage of interpretability of each factor but comes at the cost of a lower explained variance by the factor.

3.3.2. Implementation. Several solvers exist to find the solution to the constrained optimization problem in (7). In this study, we derive solutions to the problem in (7) using the augmented Lagrangian-based optimizer proposed by [26] and implemented in the open source *Rsolnp* package of [27].

4. Application

In this section, we test the impact of constraints when constructing the four pillars of the RCI. Resilience in a food security context is a concept used by the FAO to understand how households cope with shocks and stressors [23]. Currently, the FAO collaborates with organizations around the world and applies the RCI in more than ten countries in the Near East and sub-Saharan Africa [28].

The RCI is a measure of performance and is built on four fundamental attributes called pillars—namely, access to basic services (ABS), assets (AST), social safety net (SSN), and adaptive capacity (AC). Since each of these four pillars is an independent composite indicator, the application evaluates how the constraints affect loadings, attributes, and rankings of each pillar.

The remainder of this section is organized as follows. First, we introduce the concepts of the four pillars and their input variables. Subsection 4.2 describes our data sets together with their descriptive statistics, while Subsections 4.3–4.5 discuss the impact of constraints on loadings,

Table 2: Definitions of the four pillars and possible input variables.

Pillar	Input ID	Input variable description
Access to basic services (ABS): ability of a household to meet basic needs by accessing and effectively using basic services	1	Access to improved sanitation facilities
	2	Access to improved water sources
	3	Access to electricity
	4	Access to improved energy sources
	5	Closeness to primary school
	6	Closeness to hospital
	7	Closeness to markets
	8	Closeness to the nearest city
Assets (AST): key elements of a livelihood since they enable households to produce and consume goods	1	Wealth index
	2	Agricultural asset index
	3	Tropical livestock unit
	4	Land used for crop production
Social safety nets (SSN): capacity of the household to access formal and informal assistance	1	Participating in social networks
	2	Loans/credits received in the last year
	3	Formal transfers received in the last year
	4	Informal transfers received in the last year
	5	Government assistance received in the last year
Adaptive capacity (AC): ability to adapt to a new situation and develop new livelihood strategies	1	Household average years of education
	2	Share of active working members
	3	Household head can read or write
	4	Different sources of income
	5	Different crops cultivated during the last season
	6	Participating in training courses

characteristics, targeting, and correlations of pillars. Subsection 4.6 checks the stability of loading solutions.

4.1. The Four Pillars of the RCI. The four pillars, ABS, AST, SSN, and AC, are key dimensions of household resilience. Generally speaking, ABS refers to accessibility and quality of access to basic services such as clean water, improved toilets, electricity, hospitals, and schools. AST proxies productive and nonproductive assets used by the family. SSN generally includes private/public assistance and transfers received or made by the household. AC expresses the capacity to adapt after a shock and is computed using variables such as the average education of adults and the share of active working members. By its nature, each pillar is a composite indicator aggregated from a set of selected inputs. An example of these input indicators is presented in Table 2.

We see that in each subset, indicators are defined so that the higher their values, the more positive impact they have on the corresponding pillar. For example, the more land a household has, the more asset they own. By definition, the four pillars also rank better performance with higher values (e.g., an increase in assets or adaptive capacity should improve the resilience of a household against shocks and uncertainties). As such, we expect all input variables within a subset to be positively correlated with each other and with the corresponding pillars.

4.2. Descriptive Statistics. We use two data sets to test the performance of the proposed methods. The first one is collected from 4027 households living in both refugee settlements and host communities in 11 districts of Uganda in 2019. This project aims to provide a comprehensive assessment of the refugees' needs and facilitate their social and economic integration. The second data set is assembled by interviewing 6949 households in rural Chad in 2014. Its objective is to better understand the food insecurity situation under great regional heterogeneity to design effective policy responses. Since these are two projects conducted for different populations in countries with different purposes, lists of input variables used in these two sets share some similarities but are not exactly the same. For example, the ABS pillar under the Uganda 2019 project is constructed using variable IDs 1, 2, 5, 6, and 7, while the Chad 2014 data set uses variable IDs 1, 2, 3, 4, and 8 (see Table 2 for definitions of these variables).

To obtain the input indicators for the four pillars, we first gather and arrange raw information from the survey and treat the data for errors, missing observations, and outliers. Here, data points that are errors or missing are filled with median values of the local group where the household belongs, while outlying values that are more than three times the median absolute deviation from the median are replaced by the median plus three times the median absolute deviation. Then, the cleaned data are used to form required input variables. Their descriptive statistics are presented in Table 3.

To understand how input indicators interact with each other and with the overall scale, we conduct Cronbach's alpha analysis for each data subset. The Cronbach coefficient alpha [29] is the most common estimate of internal consistency of items in a survey, which assesses how well a set of individual indicators measures a single unidimensional object (e.g., attitude and phenomenon) based on their correlations [2]. Among different outputs of Cronbach's alpha analysis, we consider inter-item correlations and corrected item-total correlation results. The inter-item

Table 3: Descriptive statistics.

Pillar	Input ID	Min	Mean	Max	SD
Panel A: Uganda data set					
ABS	1	0.00	0.72	1.00	0.45
	2	0.00	0.83	1.00	0.38
	5	0.29	1.06	3.57	0.66
	6	0.02	0.16	0.83	0.14
	7	0.11	0.66	2.86	0.55
AST	1	0.00	0.46	1.05	0.18
	2	−0.46	0.27	1.36	0.29
	3	0.00	0.44	7.14	0.98
	4	0.03	1.23	7.50	1.48
SSN	1	0.00	0.51	1.00	0.50
	2	0.00	5.24	106.85	13.48
	3	0.00	4.93	25.94	5.97
	4	0.00	0.18	24.90	1.39
AC	1	0.00	6.16	15.00	3.17
	2	0.05	0.50	1.00	0.22
	4	0.00	1.48	5.00	0.89
	5	0.00	2.96	18.00	2.18
	6	0.00	0.70	1.00	0.46
Panel B: Chad data set					
ABS	1	0.00	0.41	1.00	0.49
	2	0.00	0.50	1.00	0.50
	3	0.00	0.02	1.00	0.13
	4	0.00	0.00	1.00	0.06
	8	0.00	0.18	0.44	0.10
AST	1	−1.13	−0.02	4.32	0.21
	2	−1.21	−0.02	8.82	0.32
	3	0.00	0.40	20.00	0.77
	4	0.00	0.57	55.46	2.17
SSN	2	0.00	0.49	1.00	0.50
	3	0.00	0.03	1.00	0.16
	5	0.00	0.18	1.00	0.06
AC	2	0.11	1.22	12.00	0.95
	3	0.00	0.34	1.00	0.47
	4	0.00	0.43	1.00	0.13

correlations indicate how variables interact with each other, while item-total correlations reflect how each item by itself correlated with everything else grouped together. Those statistics are presented in Table 4.

As input variables are constructed so that an increase in their values leads to an improvement in the pillar, we expect all correlations in Table 4 to be positive. However, due to the complexity of the analytical context, this assumption may be violated. In particular, in Panel A, only the AST pillar has a subset of all input indicators positively correlated. This group also has the highest values of correlations and internal consistency. In Panel B, only the SSN pillar has all positive correlations. For the other pillars, the situation is more diverse.

To better understand the negative connections between input variables, we look in detail at the ABS pillar under the Uganda project. Among five variables used to construct this composite indicator, *access to improved sanitation facilities* (input ID 1) has a negative correlation with *access to improved water source* (input ID 2) and *closeness to hospital* (input ID 6), leading to its reversed connection with the total scale. This is due to the complexity of the analytical context, as many interviewed households live in refugee settlements. These families are located far from hospitals and do not have access to clean water infrastructures that the government develops; however, they receive emergency support and mobile sanitation services provided in the camp by humanitarian projects. This complex situation creates opposing trends in responses, which leads to negative correlations between variables. Similarly, refugees receive significant formal assistance and training from governments and international organizations to help them achieve independence and self-reliance; however, they might have difficulties obtaining loans and participating in associations, or acquire limited education and sources of income. This explains the negative correlations of variables *formal transfers received* of the SSN pillar (input ID 3, Panel A) and *participating in training course* of the AC pillar (input ID 6) with the rest of their corresponding input subsets.

One way to deal with the negativity in the data set is to completely remove the opposing variables or replace them with different proxies. Doing so can increase the internal consistency of the data and ensure positive correlations among input variables and with the pillar, even under the traditional construction method. However, this approach introduces the risk of losing information. For example, it may not be possible to drop variables *access to improved sanitation facilities* and *formal transfers received*, as they capture particular features of the refugee population who receive significant investment, and policymakers want to assess how effective these budgets are spent in building resilience capacity. Thus, we propose a more robust approach of adding constraints in the pillar estimation, which provides an adjusted proxy of performance that respects the context complexity while ensuring the positive correlations between input variables and the pillar.

4.3. Impact of Constraints on Loadings and Characteristics of Pillars. In this subsection, we test the impact of constraints on loadings and characteristics of the four pillars, ABS, AST, SSN, and AC. As explained in the introduction, here we discuss in detail only the PCFA results to avoid repetition and link more with the practical usage.

Table 5 displays the results for the four pillars constructed under conventional and constrained PCFA using the Uganda 2019 and Chad 2014 data sets. First, we can see that under the traditional approach, the loadings are a mix of both positive and negative elements. Some variables such as *access to improved water sources* (input ID 2, Panel A) of the ABS pillar or *different income sources* (input ID 4, Panel A) of the AC pillar have sizable loadings in both Score 1 (positive) and Score 2 (negative), making it difficult to conclude their overall influence on the aggregate index. A meaningful interpretation of the scores is, then, not straightforward.

Notes:

(1) When employing principal component approaches to build an aggregate index, we need to decide how many intermediate scores should be retained in the analysis. Here, Kaiser's criterion [24] is applied, and

Table 4: Correlations of input indicators in Cronbach's alpha analysis.

Pillar	Input ID	Inter-item correlations					Item-total correlations
Panel A: Uganda data set							
	1	1.00					−0.16
	2	−0.11	1.00				0.18
ABS	5	0.01	0.05	1.00			0.38
	6	−0.10	0.08	0.00	1.00		0.02
	7	0.02	0.12	0.24	0.02	1.00	0.50
	1	1.00					0.57
AST	2	0.37	1.00				0.46
	3	0.27	0.19	1.00			0.47
	4	0.30	0.23	0.33	1.00		0.51
	1	1.00					0.65
SSN	2	0.20	1.00				0.43
	3	−0.09	−0.13	1.00			−0.77
	4	0.03	−0.01	−0.03	1.00		−0.02
	1	1.00					0.55
	2	0.19	1.00				0.36
AC	4	0.11	−0.01	1.00			0.16
	5	0.17	0.05	0.15	1.00		0.29
	6	−0.13	0.00	−0.18	−0.24	1.00	−0.61
Panel B: Chad data set							
	1	1.00	0.11	0.10	0.04	0.13	0.41
	2	0.11	1.00	0.04	0.03	−0.10	0.10
ABS	3	0.10	0.04	1.00	0.05	0.01	0.20
	4	0.04	0.03	0.05	1.00	0.05	0.17
	8	0.13	−0.10	0.01	0.05	1.00	0.12
	1	1.00	0.41	0.27	0.02		0.66
AST	2	0.41	1.00	0.01	0.10		0.49
	3	0.27	0.01	1.00	−0.03		0.24
	4	0.02	0.10	−0.03	1.00		0.08
	2	1.00	0.03	0.12			0.62
SSN	3	0.03	1.00	0.07			0.59
	5	0.12	0.07	1.00			0.64
	2	1.00	−0.02	0.02			0.06
AC	3	−0.02	1.00	−0.01			−0.18
	4	0.02	−0.01	1.00			0.23

we only keep scores with eigenvalues (or variance explained) above or equal to 1.0. The simplest justification for this is that it does not make sense to add a score that explains less variance than is contained in one individual indicator [2, 25]. According to this rule, no more than two scores should be retained in each of our examples. Thus, in this table, we only present estimations of the first two loading vectors given by traditional and constrained PCFA, together with correlations between input variables and pillars constructed under these approaches.

(2) A blank column indicates that the corresponding score has eigenvalues less than 1.0 and thus is not used in the construction of the pillar.

To fix this problem, we add a lower bound of zero in the optimization function to keep all loading elements positive and rotate the concentration of variable influences on scores. In particular, negative elements in the first traditional loadings (caused by the reversed effect of corresponding items with total scale; see Table 4) are converted to 0.00 under the constrained version. To ensure orthogonality between factors, values of loading elements in the second constrained component are allocated so that v_2 is perpendicular to v_1—that is, the dot product of the two vectors is zero. For example, in Panel A of Table 5, the AC pillar is constructed using five sub-indicators, among which the individual variable *participating in training courses* (input ID 6) has a negative loading in the first score under traditional approaches. With constraint, this value is converted to 0.00. The negative influence of this variable on the first score is removed and replaced with a full positive weight on the second score. The other four variables already have significant positive loadings on v_1; thus, they are no longer loaded on v_2. As such, under the constrained approach, we rotate the concentration of variable influences on scores and give straightforward interpretations for factors.

There is a slight reduction in variance explained by constrained scores compared with the traditional ones. For example, in Panel A, the first conventional factor used to construct the ABS pillar has an eigenvalue of 1.30, while its restrained version has a projected variance of 1.29. This can be explained by the nature of optimization problems used to obtain these values. The variance explained by score z_1

Table 5: Impact of constraints on loadings and pillar characteristics.

Pillar	Input ID	Conventional loadings		Correlation with pillar	Loadings with constraint		Correlation with pillar
		S1	S2		S1	S2	
Panel A: Uganda data set							
ABS	1	−0.15	0.58	0.30	0.00	1.00	0.59
	2	0.40	−0.33	0.14	0.36	0.00	0.31
	5	0.49	0.34	0.74	0.53	0.00	0.56
	6	0.20	−0.48	−0.17	0.15	0.00	0.09
	7	0.55	0.27	0.74	0.58	0.00	0.62
Variance explained		1.30	1.16		1.29	1.00	
AST	1	0.40		0.73	0.40		0.73
	2	0.35		0.64	0.35		0.64
	3	0.35		0.65	0.35		0.65
	4	0.37		0.69	0.37		0.69
Variance explained		1.85			1.85		
SSN	1	0.53	−0.03	0.52	0.65	0.00	0.58
	2	0.56	−0.18	0.46	0.64	0.00	0.55
	3	−0.42	−0.10	−0.49	0.00	1.00	0.57
	4	0.07	0.98	0.68	0.06	0.00	0.05
Variance explained		1.29	1.00		1.21	1.00	
AC	1	0.38	0.42	0.74	0.53	0.00	0.57
	2	0.18	0.72	0.69	0.35	0.00	0.43
	4	0.35	−0.31	0.22	0.35	0.00	0.31
	5	0.44	−0.12	0.46	0.47	0.00	0.41
	6	−0.42	0.31	−0.31	0.00	1.00	0.46
Variance explained		1.51	1.10		1.34	1.00	
Panel B: Chad data set							
ABS	1	0.60	0.00	0.54	0.60		0.73
	2	0.27	0.67	0.73	0.27		0.32
	3	0.42	0.16	0.49	0.42		0.51
	4	0.32	−0.07	0.23	0.32		0.39
	8	0.34	−0.66	−0.19	0.34		0.41
Variance explained		1.21	1.10		1.21		
AST	1	0.57	−0.10	0.64	0.57		0.85
	2	0.49	0.34	0.81	0.49		0.73
	3	0.31	−0.60	0.01	0.31		0.46
	4	0.11	0.67	0.55	0.11		0.16
Variance explained		1.51	1.07		1.51		
SSN	2	0.56		0.65	0.56		0.65
	3	0.39		0.45	0.39		0.45
	5	0.63		0.73	0.63		0.73
Variance explained		1.15			1.15		
AC	2	0.69	0.00	0.52	0.70	0.00	0.50
	3	−0.43	0.79	0.23	0.00	1.00	0.69
	4	0.55	0.62	0.84	0.70	0.00	0.51
Variance explained		1.03	1.00		1.02	1.00	

equals the amount of variance projected on loading vector v_1. Under the traditional approach, this number is found when solving the optimization problem (3), which has only one purpose of seeking a combination of v_1 that maximizes $v_1^T S v_1$. As such, the solution obtained here is optimized for this sole requirement and captures the highest possible variance. The constrained version in equation (7) requires the optimization problem to (a) find the maximum variance projected on this vector while (b) keeping all elements to be positive. An additional task requires a share of resources and reduces the power in explaining variance of v_1; thus, the new solution will have a lower eigenvalue than the traditional one. In general, we exchange a part of the variance captured for fixing the relation between input variables and scores. A similar logic can be applied to explain the second loading vector.

The change in loading structure leads to a shift in characteristics of the pillar. In particular, the mixed loading matrices under the classical method result in pillars being negatively correlated with one of their sub-variables (ABS, SSN, and AC in Panel A; ABS in Panel B). This indicates that an increase in input value leads to a decrease in the pillar,

Table 6: Examples of household rank shifts in the ABS pillar.

Households	Values of input variables					Values and ranks of ABS under	
	ID 1	ID 2	ID 5	ID 6	ID 7	Traditional PCFA	Constrained PCFA
A	1	1	3.33	0.04	2.50	1.52 (1)	1.11 (6)
B	1	1	3.33	0.13	2.50	1.49 (4)	1.13 (3)
C	1	1	3.33	0.14	2.50	1.48 (5)	1.14 (2)
D	1	1	3.33	0.33	2.50	1.40 (6)	1.19 (1)
E	0	1	2.00	0.20	2.00	0.64 (273)	0.29 (647)
F	0	1	3.33	0.33	1.00	0.63 (276)	0.33 (543)
G	1	0	1.00	0.33	2.00	0.41 (529)	0.32 (564)
H	1	0	0.50	0.33	2.00	0.25 (773)	0.22 (876)
I	0	0	0.33	0.13	0.14	−0.62 (4020)	−0.83 (4027)
K	0	0	0.33	0.25	0.14	−0.67 (4027)	−0.79 (4024)

which conflicts with the definitions and can dampen the representative power of the pillar. For example, the negative correlation between *closeness to hospital* (input ID 6, Panel A) and the ABS pillar suggests that staying closer to the hospital will decrease access to basic services. Under our constrained approach, the lower zero bounds keep all loading elements positive, which ensures consistent relationships between original variables, scores, and the composite indicator. All pillars constructed with constraints are positively correlated with all of their sub-indicators.

4.4. Impact of Constraints on Targeting and Rankings. The new loading structure shifts the way input variables are loaded on scores. As such, it also affects the ranking of households in each pillar. Among the pillars considered, only AST in Panel A and SSN in Panel B experience no impact under constraints, as under the original method, they use one score with only positive loading elements in the construction. Other pillars all experience changes in household performances.

To demonstrate how entity targeting shifted under constraints, we take an example of 10 families under the ABS pillar in Panel A and present their values and ranks in Table 6. These households are selected so that they display different settings from the best to the worst of the population.

Let us consider the first four households, *A*, *B*, *C*, and *D*, in Table 6, which are among the best of the examined population regarding accessibility to basic services. These households receive the same values in input variables ID 1, ID 2, ID 5, and ID 7. In other words, they all have access to clean water sources and improved toilet facilities and are equidistant from the nearest primary school and market. The only difference among these families is the distance to the closest hospital (indicated by input variable ID 6). Values that households *A*, *B*, *C*, and *D* received in indicator ID 6 are 0.04, 0.13, 0.14, and 0.33, respectively. As all input variables indicate better performances with higher numbers, this shows that household *A* stays the farthest to the hospital, second is household *B*, and so on until household *D*, who has the closest distance. Under the circumstance that all other variables are kept constant, household *D* has the easiest access to basic services; thus, we expect them to have the highest value and rank best in the ABS pillar.

However, under the traditional PCFA method, the negative loading element in the second score creates a reverse influence from input ID 6 to the composite indicator (see Table 5). Therefore, entity performance is not accurately targeted, as the values and ranks of households *A*, *B*, *C*, and *D* in the ABS pillar are contrary to their real living conditions. Here, household *A* has the best accessibility to basic services, while household *D* is ranked as the worst among the four households.

This issue is then fixed under the constrained PCFA method. The zero lower bounds ensure that all elements of the new loading matrix are positive (see Table 5). Thus, the changes in input variables are consistently reflected in values and ranks of the pillar. Household *D* is now ranked the best in accordance with their situation, while the position of household *A* is dropped to the sixth. As such, we can conclude that the constraints give a better targeting for the ABS pillar.

4.5. Impact of Constraints on Correlations between Pillars. As values and ranks of households in pillars change with new loading structures, they also affect the relationship between pillars. We can see in Table 7 that all correlations change under the impact of constraints. In some cases, the

Table 7: Correlations between pillars.

	Traditional PCFA				Constrained PCFA			
	ABS	AST	SSN	AC	ABS	AST	SSN	AC
Panel A: Uganda data set								
ABS	1.00				1.00			
AST	0.02	1.00			0.05	1.00		
SSN	0.02	0.35	1.00		0.00	0.00	1.00	
AC	0.06	0.37	0.20	1.00	0.05	0.29	−0.05	1.00
Panel B: Chad data set								
ABS	1.00				1.00			
AST	0.10	1.00			0.17	1.00		
SSN	0.12	−0.10	1.00		−0.04	−0.08	1.00	
AC	0.08	0.01	0.14	1.00	0.08	0.07	0.06	1.00

Table 8: Correlations between variables of the AST and SSN pillars for the Uganda data set.

Pillar variables		SSN			
		1	2	3	4
AST	1	0.31	0.17	−0.21	0.04
	2	0.13	0.09	−0.19	0.02
	3	0.17	0.08	−0.20	0.03
	4	0.20	0.11	−0.38	0.03

change is significant (e.g., the correlation between SSN and AST in Panel A, or between SSN and ABS under Panel B).

We analyze in detail the case of the SSN and AST pillars in Panel A to see how these changes happened. Here, the correlation drops from 0.35 under traditional PCFA to near zero under constrained PCFA. This is an important change that can have an influence on strategy development. Under the original method, we can associate more access to social safety net with more assets, meaning an increase in assistance can help a household build their possessions, which increases their resilience. This connection no longer exists under the constrained PCFA.

As loadings of the AST pillar are all positive and unchanged under the impact of constraints, this drop occurs due to the change in loadings attributed to the variable *formal transfer received* (input ID 3) of the SSN pillar. In Table 8, we see that *formal transfer received* has significant negative correlations with all variables under the AST pillar. These numbers reflect the fact that the more assets a household has, the less likely they are to receive support from governments and humanitarian organizations. Then, combined with the negative loadings attributed to the variable *formal transfer received* under the traditional construction approach (see Table 5), these negative correlations are reversed, resulting in a positive link between the SSN and AST pillars.

In contrast, under the constrained approach, all loadings are kept positive. Thus, the negative connections between *formal transfer received* and AST variables are reflected as they are to the pillar. These adverse effects are then counterbalanced by the positive relations between the remaining SSN variables and the AST variables. As such, the new correlation of SSN and AST is near zero, as presented in Table 7.

Table 9: Mean of absolute loading difference.

Pillar	PCFA	Constrained PCFA
Panel A: Uganda data set		
ABS	0.24	0.05
AST	0.01	0.01
SSN	0.18	0.03
AC	0.12	0.04
Panel B: Chad data set		
ABS	1.04	0.14
AST	0.22	0.08
SSN	0.04	0.04
AC	0.67	0.07

4.6. Stability of v_k. To see how stable the loading solutions are, we conduct a stability test using mean absolute error (MAE) as the criterion. First, we calculate the loading matrix using the complete data set as the comparison base. Next, a new loading solution is estimated using a randomly drawn subsample that covers 75% of the population. We subtract these matrices to obtain the absolute difference. This is repeated 500 times with different subsamples. Then, the MAE is defined as follows:

$$\text{MAE} = \frac{1}{500}\sum_{i=1}^{500}\left|V^{ba} - V_i^{su}\right|, \tag{8}$$

where V^{ba} is the base loading solution calculated using the full data set and V_i^{su} is the i^{th} new loading solution estimated using the randomly drawn subsample ($i = 1, \ldots, 500$). We apply this test for 16 aggregate indicators (formed by 4 pillars × 2 construction methods × 2 data sets) and summarize their MAE in Table 9.

Overall, we can see that constrained PCFA returns loading solutions that are more stable than the traditional approach. In most cases, MAEs of pillars constructed under

constrained PCFA are much lower compared with their versions under standard PCFA. Exceptions are AST in Panel A and SSN in Panel B: the lower bounds do not change the loading structures of these pillars; thus, their MAEs are the same under the two construction methods.

5. Conclusions

In this study, we document how adding constraints improves the representability of principal component-based composite indicators. The traditional approaches use loading solutions that include both positive and negative values. This introduces potential difficulties in interpretations, inaccurate ranking, and conflicts with the theoretical framework for the composite indicator. To fix these issues, we propose to apply a constraint in the process of finding each loading vector. This restriction ensures a consistent relation between input variables, the scores, and the aggregate index.

We compare the performance of conventional and constrained principal component methods in constructing the four pillars of the RCI. This is an approach used by the FAO to understand how households cope with shocks and stressors in the context of food security. Using two data sets collected from Uganda in 2019 and from Chad in 2014, we see that the constraints have a material impact on the interpretability of the four pillars. The restricted loading matrix ensures positive relations between pillars and their input variables as guided in the resilience theoretical framework. The new matrix also influences household targeting and the connections between pillars, which is of interest to policymakers. We conclude that the robustness gain of using the constrained method strikes a balance between the objective of using a small number of factors with high explanatory power, on the one hand, and the interpretability of the obtained composite indicator, on the other hand.

Data Availability

The empirical application involves data collected by FAO and UN to support the development of different communities in Uganda and Chad. These datasets can be obtained from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] K. Boudt, V. Todorov, and W. Wang, "Robust distribution-based winsorization in composite indicators construction," *Social Indicators Research*, vol. 149, no. 2, pp. 375–397, 2020.

[2] OECD, *Handbook on Constructing Composite Indicators: Methodology and User Guide*, OECD Publishing, Paris, France, 2008.

[3] G. Nicoletti, S. Scarpetta, and O. Boylaud, "Summary indicators of product market regulation with an extension to employment protection legislation," *OECD, Economics Department Working Papers*, 2000.

[4] T. Li, H. Zhang, C. Yuan, Z. Liu, and C. Fan, "A PCA-based method for construction of composite sustainability indicators," *International Journal of Life Cycle Assessment*, vol. 17, no. 5, pp. 593–603, 2012.

[5] D. Filmer and L. H. Pritchett, "Estimating wealth effects without expenditure data — or tears: an application to educational enrollments in states of India," *Demography*, vol. 38, no. 1, pp. 115–132, 2001.

[6] S. O. Rutstein, *The DHS Wealth Index: Approaches for Rural and Urban Areas*, Macro International, Calverton, MD, USA, 2008.

[7] MICS, "Palestinian multiple indicator cluster survey 2019–2020," *Technical Report, United Nations Children's Emergency Fund*, 2020.

[8] R. Reris and J. P. Brooks, "Principal component analysis and optimization: a tutorial," in *14th informs computing society conference, Richmond, Virginia*, 2015.

[9] J. Krishnakumar and A. L. Nagar, "On exact statistical properties of multidimensional indices based on principal components, factor analysis, MIMIC and structural equation models," *Social Indicators Research*, vol. 86, no. 3, pp. 481–496, 2008.

[10] Penn State, "STAT 505: applied multivariate statistical analysis - 12.12-Estimation of factor scores," The Pennsylvania State University, State College, PA, USA, 2021.

[11] M. Hubert, P. J. Rousseeuw, and S. Van Aelst, "Multivariate outlier detection and robustness," *Handbook of Statistics*, vol. 24, pp. 263–302, 2005.

[12] P. J. Rousseeuw, M. Debruyne, S. Engelen, and M. Hubert, "Robustness and outlier detection in chemometrics," *Critical Reviews in Analytical Chemistry*, vol. 36, no. 3–4, pp. 221–242, 2006.

[13] K. A. Bollen, "Outliers and improper solutions: a confirmatory factor analysis example," *Sociological Methods & Research*, vol. 15, no. 4, pp. 375–384, 1987.

[14] S. Engelen and M. Hubert, "Detecting outlying samples in a parallel factor analysis model," *Analytica Chimica Acta*, vol. 705, no. 1–2, pp. 155–165, 2011.

[15] G. Pison, P. J. Rousseeuw, P. Filzmoser, and C. Croux, "Robust factor analysis," *Journal of Multivariate Analysis*, vol. 84, no. 1, pp. 145–172, 2003.

[16] J. Choi, G. Oehlert, and H. Zou, "A penalized maximum likelihood approach to sparse factor analysis," *Statistics and Its Interface*, vol. 3, no. 4, pp. 429–436, 2010.

[17] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of Computational & Graphical Statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[18] G. I. Allen, L. Grosenick, and J. Taylor, "A generalized least-square matrix decomposition," *Journal of the American Statistical Association*, vol. 109, no. 505, pp. 145–159, 2014.

[19] T. Ozeki, K. Koide, and T. Kimoto, "Evaluation of sources of acidity in rainwater using a constrained oblique rotational factor analysis," *Environmental Science and Technology*, vol. 29, no. 6, pp. 1638–1645, 1995.

[20] P. Paatero, "Least squares formulation of robust non-negative factor analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 37, no. 1, pp. 23–35, 1997.

[21] P. Paatero and U. Tapper, "Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

[22] S. Lipovetsky, "PCA and SVD with nonnegative loadings," *Pattern Recognition*, vol. 42, no. 1, pp. 68–76, 2009.

[23] FAO, "RIMA-II — resilience index measurement and analysis—II," *Technical report, Food and Agriculture Organization of the United Nations*, FAO, Rome, Italy, 2016.

[24] H. F. Kaiser, "The application of electronic computers to factor analysis," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 141–151, 1960.
[25] I. T. Joliffe, *Principal Component Analysis*, Springer, Berlin, Germany, 2nd edition, 2002.
[26] Y. Ye, *Interior algorithms for linear, quadratic, and linearly constrained non-linear programming*, Ph.D. Thesis, Stanford University, Stanford, CA, USA, 1987.
[27] A. Ghalanos and S. Theussl, "Rsolnp: general non-linear optimization using augmented Lagrange multiplier method," *R package version*, vol. 1.16, 2015.
[28] FAO, "Resilience index measurement and analysis (RIMA)," *Web page, Food and Agriculture Organization of the United Nations*, FAO, Rome, Italy, 2021.
[29] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," *Psychometrika*, vol. 16, no. 3, pp. 297–334, 1951.