Published in final edited form as:
Ann Epidemiol. 2018 October ; 28(10): 736–738. doi:10.1016/j.annepidem.2018.08.003.

On a scale as a sum of manifest variables

Hee-Choon Shin, PhD[*]
National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD

Abstract

The most common approach for a scale construction is to create a scale as a sum of manifest variables (a "sum scale"). When we use the sum scale for analysis, we implicitly assume that there is a one-dimensional latent structure representing the manifest data on a multidimensional space. In this commentary, we review basics of identifying a latent structure using measured variables with a minimum linear algebra. We demonstrate the technique using Fisher's iris data as an illustration. We examine the relationships between resulting latent variables and the sum scale to evaluate goodness of the sum scale. As a practical solution, in general, we could create a sum scale using a set of positively and highly correlated measured variables. More care is needed when the data are not unidimensional.

Keywords

Latent structure; Scale; Singular value decomposition (SVD); Eigenvalue; Eigenvector; Space; Subspace; Dimension

Introduction

In epidemiological studies, we indirectly measure most of research variables (e.g., health, disability, or cardiovascular fitness) with manifest variables because we cannot directly measure the latent constructs or variables. We can think of the manifest variables as proxy measures for the true measure. Frequently, we use a scale to represent the hidden or latent construct based on the manifest variables. The most common approach is to create a scale as a sum of measured variables (a "sum scale"). This short note is to describe some important considerations for scale construction with manifest measures after data collection, but we need to note that a careful design and implementation is necessary for a quality scale before data collection [1,2]. Indeed, the whole process of scale construction is too big a topic for this short commentary.

Let us consider P manifest measures for each of n subjects or a n × P rectangular data matrix A. Typically n is much larger than P. Geometrically, we can visualize n points on a P-

[*]Corresponding author. National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782. Tel.: +1-301-458-4307. hshin@cdc.gov.

Conflict of interest statement: H-C. Shin has no conflicts of interest.

Disclosure statement: H-C. Shin has no financial disclosures.

The findings and conclusions stated in this commentary are solely those of the author. They do not necessarily reflect the views of the National Center for Health Statistics or the Centers for Disease Control and Prevention.

dimensional space (row picture) or P vectors on n-dimensional space (column picture). Now our objective is to identify a one-dimensional subspace in the P-dimensional space summarizing the data and to derive a single latent variable on the one-dimensional subspace. As an extreme but simple example, consider a set of data points exactly on the 45° line on a 2-dimensional plane of manifest variables. We would not need two variables to describe these data. We can reduce the 2-dimensional space to a one-dimensional subspace and describe the data on the 45° line. Now just a number is sufficient to describe a two-dimensional data point. We can extend this logic to nth dimensional space.

The “Sum” scale

The sum scale is simply defined as an unweighted sum of P manifest measures. For the current discussion, we assume that there is no item missing values. The sum is scale dependent. Therefore, each item needs to be standardized before being summed if units of measurement are not uniform. A weighted sum scale could be constructed by giving unequal weight to each item depending on the importance of each item toward the sum scale. This sum scale is simple and easy to construct but is being utilized without any rigorous mathematical justification.

Unidimensionality of the latent structure

When we create a scale as a sum of P variables, we implicitly assume that there is a one-dimensional hidden structure. Therefore, the most important criterion for a sum scale is the unidimensionality of A. The dimensionality can be evaluated by a technique called singular value decomposition (SVD), which is similar to principal components analysis [3], and many of free or commercial software are available for SVD (e.g., a R or MATLAB function). Details on the SVD and other topics (including eigenvalues and eigenvectors) of linear algebra can be found in any standard linear algebra textbook [4]. According to SVD technique, any n × P matrix A can be a multiplication of the following three matrices: n × P orthogonal matrix U = [u_1, u_2 … u_p], a P × P diagonal matrix $\sum$ of singular values (σ_i), and a P × P orthogonal matrix, V^T = [v_1, v_2 … v_p]^T, where T stands for a transpose. That is, A = $U\sum V^T$ or $A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \ldots + \sigma_r u_r v_r^T$. As orthogonal matrices, $U^TU = V^TV = I$, where I stands for a P × P identity matrix with 1’s at its diagonal and 0’s for off-diagonal. A square of the singular value is its eigenvalue, and there could be P–r zero singular values. The ratio (λ_1) of the first and largest eigenvalue $\left(\sigma_1^2\right)$ to the sum of all eigenvalues would tell us the unidimensionality of the n × P Pmatrix A. The λ_1 should be large (e.g., 0.8) to claim a unidimensionality of the latent structure. The value of 0.8 indicates that 80% of variation of the data is explained by the first component ($\sigma_1 u_1$). A larger λ_1 implies that the configuration on the P-dimensional space can be fairly represented by the new configuration on the one-dimensional subspace, embedded in the P-dimensional space. When λ_1 is large, we can use the first component, a new variable ($Av_1 = \sigma_1 u_1$), as a latent construct for analysis instead of the P manifest variables. The new latent variable is the best approximation of the multivariate data as a linear combination of the P manifest variables. Unidimensionality is necessary but not sufficient for a good sum scale. Soundness of the sum scale can be evaluated by looking at the relationship between the first component and the sum scale.

Multidimensionality of the latent structure

When λ_1 small (e.g., 0.3), the latent structure is multidimensional. When λ_1 is relatively small, each component should be analyzed separately because each component measures a different latent variable. When theory permits, however, we could create an overall scale or index by combining all the latent components. For example, various study subjects' scores can be summed to measure an overall achievement of a student. Here, we assume that each component additively contribute to the overall latent measure. That is, we could create an overall scale by summing all latent variables ($\sigma_1 u_1$, $\sigma_2 u_2$ … $\sigma_r u_r$). However, it is not recommended without specific theoretical rationale. Adding various subjects' test components to measure the overall achievement for a student may be conceptually sound, but adding physical and mental disability components for an overall disability scale may not be. Utilizing singular values and singular vectors, the single overall latent variable (ξ) is $\sigma_1 u_1 + \sigma_2 u_2 + \ldots + \sigma_r u_r$, which is a weighted sum of singular vectors, the weights being singular values. When the latent structure is multidimensional, ξ should be better than the sum scale. An important consideration is that all the components need to be identified first before being summed to create the scale. Soundness of the sum scale could be evaluated by looking at the relationship between ξ and the sum scale.

A demonstration using Fisher's iris data (1936)

Fisher (1936) observed four variables/characteristics (sepal length, sepal width, petal length, and petal width) of 50 plants in each type of three irises (iris setosa, iris versicolor, and iris virginica) [5]. Overall, there were 150 units and there were no missing values. Here, we are not concerned with the nature of the four measures and we are simply interested in the numerical values as a given set of data. All the four variables were measured in cm and therefore the raw measures can be analyzed without incurring any unit of measurement issues. If units of measurement are different, we may need to rescale or standardize the measures because SVD is scale dependent. Table 1 shows correlation coefficients among the four manifest variables and other latent variables. Sepal width is negatively related to the other three manifest variables, but the correlation coefficients among the other three manifest variables ranges from 0.8179 to 0.9629. The eigenvalues are 9208.31, 315.45, 11.98, and 3.55. The four eigenvalues are unique and larger than zero, which indicate that the four manifest variables are linearly independent [6]. The λ_1 is 0.9653, which indicates that 96.53% of the variation of the data on the 4-dimensional space is explained by the first component. We can safely conclude that the latent structure is one-dimensional and use the first component ($\sigma_1 u_1$) for analysis. The first component is negatively related to sepal width (−0.2201) but is highly correlated with other three manifest variables, indicating that relatively sepal width is not well represented by the first component. In fact, the third component is moderately related to the sepal width (0.5241).

Even if one-dimensionality is confirmed, however, the sum scale may not be the valid scale if the first component and the sum scale are not related. We examine the relationship between the first component and the scale as a sum of the four manifest variables. It happens that the relationship is almost perfect. The correlation coefficient is 0.9968. Results of using either the first component or the sum measure should be similar.

We can construct a scale as a sum of latent components when the latent structure is multidimensional. The latent construct ξ could be constructed as a sum of the four components: $\xi = \sigma_1 u_1 + \sigma_2 u_2 + \sigma_3 u_3 + \sigma_4 u_4$. The correlation coefficients with the first component and the sum scale are 0.9719 and 0.9839, respectively, which indicates soundness of the sum scale. Note that the current data (one-dimensional) are not ideal to evaluate the relationship between ξ and the sum scale.

If theory permits, we could omit the sepal width that is negatively related to all the other three measures and conduct the analysis for a better first component. When we omit the sepal width, λ_1 is 0.9737 (See Table 2), which is larger than the one (0.9653) with four manifest variables. We could use the three items omitting sepal width for a better first component. If sepal width is an essential item in constructing a scale based on a certain theory, however, sepal width should not be omitted in the analysis even if the item is negatively related to the other three items. Table 3 shows correlation coefficients of latent constructs between when the all four variables are used and when we omit sepal width. As we see in Table 3, they are highly correlated. Practically, any one of the six latent constructs could be used as a scale. We should note the data that are unidimensional and consequently omitting a variable (sepal width) do not change the three scales ($\sigma_1 u_1$, ξ, and SUM) in a significant way.

Summary

For this particular example using the Fisher's iris data, we can conclude that the sum scale (SUM) along with the other two latent scales ($\sigma_1 u_1$ and ξ) is a good measure for the latent construct. However, we should keep in mind that it might not be the case for other data. Consequently, we need a careful analysis before creating a sum scale as a hidden construct. At a minimum, we need to look at the relationships among the manifest variables whether the manifest variables measure the same hidden structure. If the latent structure is one-dimensional, all the absolute values of correlation coefficients among the manifest variables should be large because each manifest variable measures the same hidden structure. Because of various reasons (e.g., measurement error), there could be some items that are not related or negatively related to the other items. Those items could be omitted for a sum scale if theory permits. However, we always need to keep in mind that the initial assumption of unidimensionality could be wrong. Algebraically, the first component ($\sigma_1 u_1$) is the best measure for the hidden one-dimensional construct. As a practical alternative, we could create a sum scale using only positively and highly correlated variables that would generate a larger λ_1.

Categorical and ordinal variables

Frequently, the manifest variables are measured in categorical or ordinal terms. In many cases, continuous measures are transformed into categorical variables (one for presence of certain characteristics, zero for the absence; one for high, zero for low; one for low, two for medium, three for high; etc.) and then summed to create a scale. As long as we accept the numerical values as integer scores, all the discussion in the above are applicable. If one hypothesize a categorical latent structure or a latent class (e.g., healthy or not healthy) instead of a continuous scale, other techniques such as latent class analysis can be applied to the data [7–10].

References

[1]. Comrey AL. Factor-analytic methods of scale development in personality and clinical psychology. J Consult Clin Psychol 1988;56(5):754–61. [PubMed: 3057010]

[2]. Dawis RV. Scale construction. J Couns Psychol 1987;34(4):481–9.

[3]. Kendall M. Multivariate analysis. New York: Hafner Press; 1975.

[4]. Strang G. Introduction to linear algebra. 4th ed Wellesley, MA: Wellesley-Cambridge Press; 2009.

[5]. Fisher RA. The use of multiple measurements in taxonomic problems. Ann Eugen 1936;7:179–88.

[6]. Rogers JL, Nicewander WA, Toothaker L. Linearly independent, orthogonal, and uncorrelated variables. Am Stat 1984;38(2):133–4.

[7]. Agresti A. Categorical data analysis. 2nd ed. New York: Wiley; 2002.

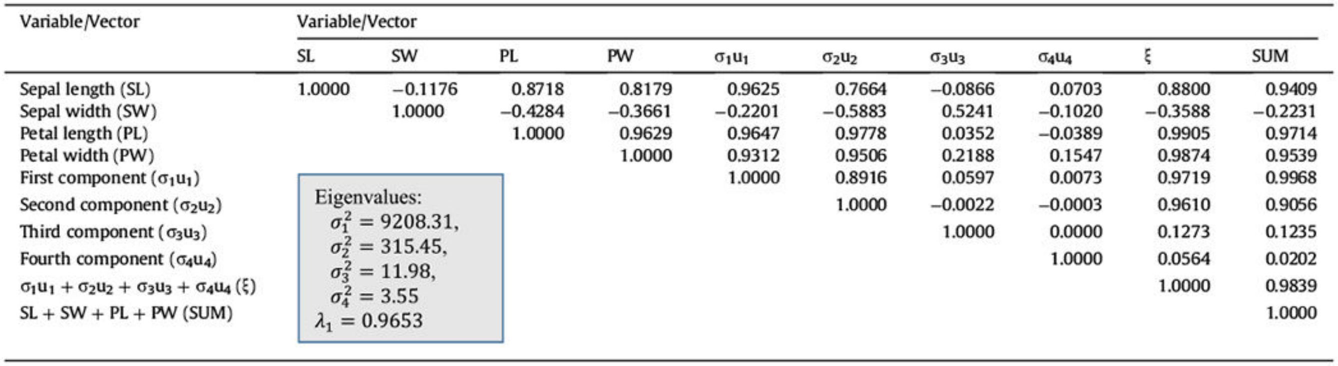
[8]. Clogg CC. Some latent structure models for the analysis of likert-type data. Soc Sci Res 1979;8:287–301.

[9]. Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika 1974;61(2):215–31.

[10]. McCutcheon A. Latent class analysis. Newbury Park: Sage; 1987.

Table 1

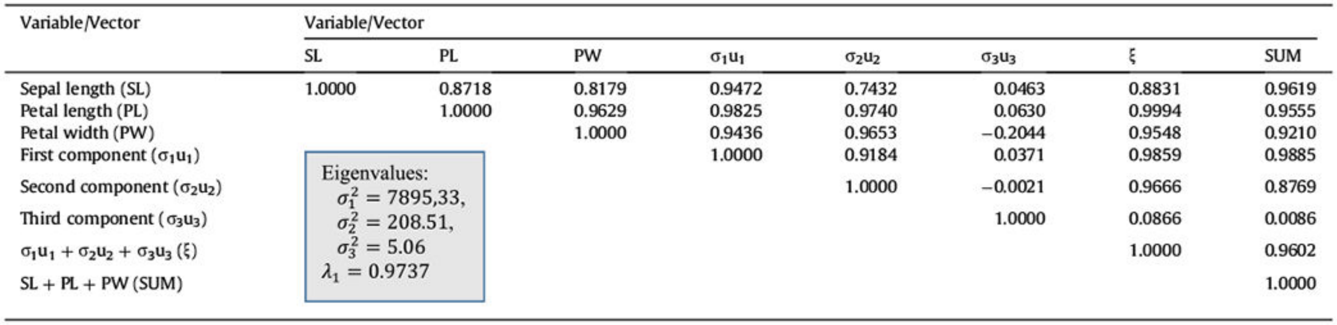
Correlation coefficients: Fisher’s iris data (1936) and derived latent variables



Variable/Vector	Variable/Vector									
	SL	SW	PL	PW	$\sigma_1 u_1$	$\sigma_2 u_2$	$\sigma_3 u_3$	$\sigma_4 u_4$	ξ	SUM
Sepal length (SL)	1.0000	−0.1176	0.8718	0.8179	0.9625	0.7664	−0.0866	0.0703	0.8800	0.9409
Sepal width (SW)		1.0000	−0.4284	−0.3661	−0.2201	−0.5883	0.5241	−0.1020	−0.3588	−0.2231
Petal length (PL)			1.0000	0.9629	0.9647	0.9778	0.0352	−0.0389	0.9905	0.9714
Petal width (PW)				1.0000	0.9312	0.9506	0.2188	0.1547	0.9874	0.9539
First component ($\sigma_1 u_1$)	Eigenvalues: $\sigma_1^2 = 9208.31$, $\sigma_2^2 = 315.45$, $\sigma_3^2 = 11.98$, $\sigma_4^2 = 3.55$ $\lambda_1 = 0.9653$				1.0000	0.8916	0.0597	0.0073	0.9719	0.9968
Second component ($\sigma_2 u_2$)						1.0000	−0.0022	−0.0003	0.9610	0.9056
Third component ($\sigma_3 u_3$)							1.0000	0.0000	0.1273	0.1235
Fourth component ($\sigma_4 u_4$)								1.0000	0.0564	0.0202
$\sigma_1 u_1 + \sigma_2 u_2 + \sigma_3 u_3 + \sigma_4 u_4$ (ξ)									1.0000	0.9839
SL + SW + PL + PW (SUM)										1.0000

Table 2

Correlation coefficients: Fisher’s iris data (1936) without sepal width and derived latent variables



Variable/Vector	Variable/Vector							
	SL	PL	PW	$\sigma_1 u_1$	$\sigma_2 u_2$	$\sigma_3 u_3$	ξ	SUM
Sepal length (SL)	1.0000	0.8718	0.8179	0.9472	0.7432	0.0463	0.8831	0.9619
Petal length (PL)		1.0000	0.9629	0.9825	0.9740	0.0630	0.9994	0.9555
Petal width (PW)			1.0000	0.9436	0.9653	−0.2044	0.9548	0.9210
First component ($\sigma_1 u_1$)	Eigenvalues: $\sigma_1^2 = 7895,33,$ $\sigma_2^2 = 208.51,$ $\sigma_3^2 = 5.06$ $\lambda_1 = 0.9737$			1.0000	0.9184	0.0371	0.9859	0.9885
Second component ($\sigma_2 u_2$)					1.0000	−0.0021	0.9666	0.8769
Third component ($\sigma_3 u_3$)						1.0000	0.0866	0.0086
$\sigma_1 u_1 + \sigma_2 u_2 + \sigma_3 u_3$ (ξ)							1.0000	0.9602
SL + PL + PW (SUM)								1.0000

Table 3

Correlation coefficients among latent constructs

		With three items (without sepal width)		
		$\sigma_1 u_1$	**ξ**	**SUM**
With four items	First component ($\sigma_2 u_1$)	0.9948	0.9689	0.9985
	Sum of latent components (ξ)	0.9822	0.9875	0.9640
	Sum scale (SUM)	0.9921	0.9730	0.9955