

Partial Least Squares and Principal Component Analysis with Non-metric Variables for Composite Indices

Jisu Yoon
Abstract

A composite index is an aggregated variable comprising individual indicators and weights that commonly represent the relative importance of each indicator. Composite indices are often used to measure latent phenomena or to summarize complex information in a small number of variables. It is crucial to choose correct weights for the variables that build a composite index. Principal Component Analysis (PCA) is a popular approach to derive weights, but it may not work when informative variations account for only small variances in the variables in a composite index. Therefore, this study proposes to use Partial Least Squares (PLS), which takes advantages of the relationship between outcome variables and the variables in a composite index. Our simulation study shows that PLS performs either as good as PCA or significantly outperforms it. Additionally, in practice variables that enter a composite index are often non-metric, which require special treatments to apply PCA or PLS. This study reviews various PCA and PLS algorithms for non-metric variables available in the literature and compares them by means of extensive simulation studies to make recommendations for practitioners. Dummy coding shows often satisfactory performance compared to more sophisticated methods. As our applications wealth, globalization, gender inequality and corruption are quantified using composite indices based on PCA and PLS, by which PLS generates composite indices tailored to each respective outcome variable showing often better performance compared to PCA. A comparison between PCA and PLS weights and coefficients shows which variables are particularly relevant for each respective outcome variable.

Ein zusammengesetzter Index ist eine aggregierte Variable, die aus individuellen Indikatoren und Gewichten besteht, wobei die Gewichte die relative Wichtigkeit jedes Indikators darstellen. Zusammengesetzte Indizes werden oft benutzt um latente Phänomene zu schreiben oder komplexe Informationen zu einer geringen Anzahl an Variablen zusammenzufassen. Es ist von großer Bedeutung richtige Gewichte für die Variablen, die einen zusammengesetzten Index bilden, zu wählen. Hauptkomponentenanalyse (PCA) ist ein populärer Ansatz um Gewichte abzuleiten, aber es ist ungeeignet, wenn informative Variationen nur kleine Varianzen der Variablen in einem zusammengesetzten Index haben. Deshalb schlägt diese Studie vor, Partial Least Squares (PLS) anzuwenden, welches die Beziehung zwischen Zielvariablen and den Variablen in einem zusammengesetzten Index ausnutzt. Unsere Simulationsstudie zeigt, dass PLS so gut wie PCA funktioniert oder erheblich es übertrifft. Zusätzlich sind in der Praxis die Variablen in einem zusammengesetzten Index häufig nicht-metrisch. Solche Variablen benötigen spezielle Verfahren, um PCA oder PLS anzuwenden. Diese Studie untersucht mehrere PCA und PLS Algorithmen für nicht-metrische Variablen in der vorliegenden Literatur und vergleicht sie durch umfangreiche Simulationsstudien, um Empfehlungen für die Praxis abzugeben. Dummy coding zeigt häufig zufriedenstellende Leistung im Vergleich zu komplizierteren Methoden. Als unsere Anwendungen betrachten wir Vermögen, Globalisierung, Geschlechtergleichheit und Korruption, indem PCA- und PLS-basierte

zusammengesetzte Indizes angewendet werden. PLS erzeugt für die jeweiligen Zielvariablen massgeschnittene zusammengesetzte Indizes, die häufig bessere Leistung als PCA zeigten. Ein Vergleich zwischen PCA und PLS Gewichten und Koeffizienten zeigt, welche Variablen für die jeweiligen Zielvariablen besonders relevant sind.