Index construction and multivariate analysis in high-dimensional environments: Application to a cultural policy index

Andrej Srakar
Institute for Economic Research, Ljubljana and Faculty of Economics, University of Ljubljana, Slovenia

Vesna Čopič
Faculty of Social Sciences, University of Ljubljana

The new era of statistics has been described as the »age of big data« (Hellerstein 2008). Although describing a different problem, problems of high-dimensional data have also been ever more frequently present in statistical and econometric literature in past years (see e.g. Chernozhukov & Hansen 2013; Belloni, Chernozhukov & Hansen 2011; Koch 2013), describing statistical analysis in the presence of much higher number of variables than observations. In our article we will analyze the problems of high-dimensionality when constructing indexes for a selected array of data. Such problems frequently abound when using macroeconomic data when data exist for only a selected group of countries (usually a smaller number) and one wants to estimate a composite measure (index) for benchmarking and statistical comparisons. Tools from multivariate analysis (factor analysis, principal components, clustering, multidimensional scaling) can be fruitfully used to this purpose as well as shrinkage and penalized regressions (see e.g. Fan & Lv 2008; Fan & Lv 2010; Fan et al. 2010).

In the first part of the article we will therefore present the problem in a general statistical framework showing that high-dimensionality of data can significantly alter the construction of an index, particularly in sparse environments. We will present some new statistical results and solutions for the robustness of the constructed index. In the second part we will apply the findings to the construction of a »cultural policy index«, based on data for over 400 variables for 35 European countries from Eurostat's Cultural Statistics 2007 and 2011 Pocketbooks, Eurostat and Compendium for Cultural Policies and Trends in Europe. As cultural statistics is characterized by problems of unclear classification and missing data, imputation methods have to be used, as well as methods from high-dimensional multivariate and regression analysis (see e.g. Koch 2013). The results will enable us to present a new measure in the field of cultural policy and cultural statistics to estimate the performance of cultural policy of a certain country (at present almost no such tools based on solid statistical analysis exist). Generalizations to indexes in other sectors in the economy and the public sector will also be made.