Final Project in Computational Statistics – Outline

Marc Lipfert
(Matriculation Number: 3220513)

July 14, 2020

Principal Component Analysis

The overarching topic of my final project is Principal Component Analysis (PCA). Therefore, the first part of my notebook will be devoted to introducing the method. In particular, I will outline its usefulness with respect to dimensionality reduction.

As a second step, I will propose a context in which PCA can be applied. Over the last two decades, economic research has increasingly recognised the importance of personality traits for labour market outcomes. There is an extensive literature that emphasises the role of a particular non-cognitive trait: locus of control (LOC). It captures the extent to which an individual believes that her life can be shaped by her own actions and decisions (internal LOC) or is instead contingent on outside factors beyond her control (external LOC). The latter sentiment, i.e. the conviction that one's course of life depends on fate or luck, has been found to be detrimental for labour market success. Measures of LOC are usually elicited using multiple survey items which are designed to cover different facets of LOC. Each item is usually scored on a Likert scale.

In the early literature, researchers then simply took the (standardised) average of the item scores as their measure of LOC. For instance, Heineck and Anger (2010) study the returns to cognitive and non-cognitive personality traits using survey data from the German Socio-Economic Panel (SOEP). The SOEP comprises 10 items for LOC, each of them scored on a 10-point Likert scale. The authors then take a standardized average and obtain a significantly negative effect of external LOC on wages.

In a simulation study, I will compare this approach of simply averaging item scores with PCA. The setting from the SOEP serves a starting point with respect to the number of items and other relevant aspects. The appropriateness of the respective approaches is evaluated based on the predictive performance (Mean Squared Error). If time and space constraints permit, I will additionally perform this analysis using actual data from the SOEP in place of simulated data.