Methods and Initial Results

Ken Chen

Part 1. Models and Methods

PCA to coin measures of economic hierarchy

As stated in my literature reviews, previous studies on intergenerational mobility have placed their focus on absolute terms of personal income, especially labor incomes. However, these metrics do not form a comprehensive outlook of one's economic well-being. Taken more dimensions of income variables into account, we might also be concerned about that what differentiated the economic well-being among the parents and the children might not be the same. Therefore, I'd like to urge the use of Principal Component Analysis to extract the principal component of the 'high-dimensional' income data, which can help to construct our measure of the economic hierarchy.

Our income metric consists of several subcategories:

$$A = (Labor_{inc} \quad SOB_{inc} \quad Property_{inc} \quad Labor_{inc})$$

To utilize the PCA component, we determine to use the first principle component as the measure of economic status. In other words, we are solving the following optimization problem:

$$\arg\max_{x} Var(Ax)$$

Then, by projecting each individual's income vector onto $\boldsymbol{x}$, we can obtain our measure. This is going to be done separately for the parents and children. Let us name the measure as 'economic score'.

Predicting models

This part serves for the purpose of feature selection. The quantitative component of this project is primarily about measuring intergenerational income elasticity and how much of that is realized via the channel of education investment. Thus, we proposed two models to see detect such effects.

- First, we'll define a random forest classifier, including variables of education and parents' income, to see how well these factors can help predict how economically accomplished the children can be. We turned the question into a classification problem, by dividing the parents and children into five layers based on their score on our measure.

- Second, we will introduce a more econometrics-prone method to measure the intergenerational income elasticity:

$$e_{p\to c} = \frac{\partial ln(Inc_{children})}{\partial ln(Inc_{parents})}$$

To prohibit the effects of noisy predictors, we espouse the use of LASSO models to perform the feature selection. This model could even be broken down by the parents' working profile. Interesting findings might be obtained from studying different industries and occupations.

Part 2. Data

My data is a survey dataset, named China Household Survey Data, which is collected by the National Bureau of Statistics of China. Until 2006, the survey was initiated in 1986, and has been conducted for 5 rounds. It covered most of the country's provinces and cities. The samples were drawn through a multi-step selection procedure. The survey provides very detailed data on income and expenditure of the urban households in China. In fact, the dataset follows a rotating panel structure, with one-third of samples updated each year, and each new round will take an entirely new sample. One prominent issue with the dataset is that different rounds of surveys are collecting slightly different variables, and the variable labels are not consistent. For this research project, due to the constraint of time, we plan to use only the fifth round of data (from 2002 to 2006).

The data can only be accessed once submitting access to the National Bureau of Statistics of China. I obtained the data in Apr. 2018. For each year, there are two tables. One of them is an individual based table, while the other is a household-based table. We merged the two kinds of tables, on the individual identifying index, so that people from the same household will share identical family variables. One major issue with the data is that children might leave the house to form a household entity when they are wealthy enough, and those who remain in the household might not be as economically successful as those who leave. Therefore, we filtered the merged table to contain only those who have just entered the workforce, by bounding the children's age between 16 and 30, and the parents' income variables were spread to become new columns of the table. We care most about the linkage between parents' income and children's income at the early phases of their careers.

Part 3. Initial Results

Descriptive Statistics

	total	wage	other labor	sob	property	transfer	p_total	p_wage	p_other labc	p_sob	p_property	p_transfer
count	30486	30486	30486	30486	30486	30486	30486	30486	30486	30486	30486	30486
mean	8832.56856	7941.47308	184.722262	347.446952	16.7971643	342.1291	14843.7058	9491.60858	291.214693	669.395212	260.903053	4130.5843
std	9803.20279	9305.1907	987.142899	3257.8503	377.028099	2254.24755	12481.8592	11899.8124	1710.8467	4881.56545	1933.66199	7034.66454
min	2	0	0	0	0	0	0	0	0	0	0	0
25%	2730	900	0	0	0	0	7388	0	0	0	0	0
50%	6473.8999	5960	0	0	0	0	12003.1001	6541	0	0	0	240
75%	11729.3499	11184	0	0	0	0	18420	14508.0524	0	0	0	6940
max	307300	142200	45600	307300	50000	190000	298720	159250	117355	298600	144000	147007.203

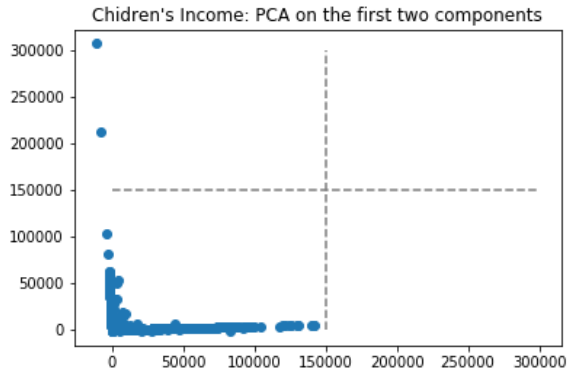
Here we present the first table, which is the descriptive statistics of the income information, broken down by subcategories. As we can see, the wage earnings formed the primary part of the income. This is true for both the parents and the children. One important observation from the table is that children's wage earnings are quite comparable to the parents'. But this property does not extend to total income, probably due to the fact that parents have more other income sources, and these other incomes are marked with very different structure between the parents and children. Therefore, it might be unfair to use the original earnings data as a measure of economic status. Rather, we would request for a within-group measure and normalize their 'economic score' so that they can be comparable on the same scale.

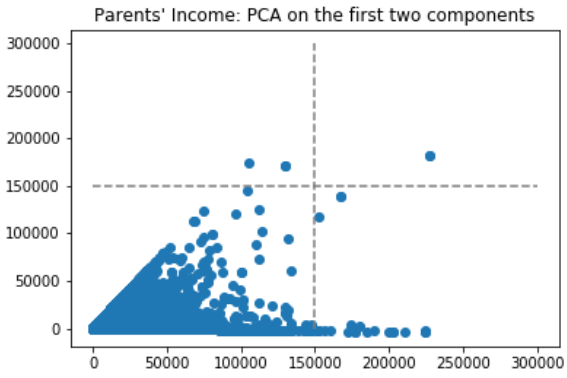
Principal Component Analysis

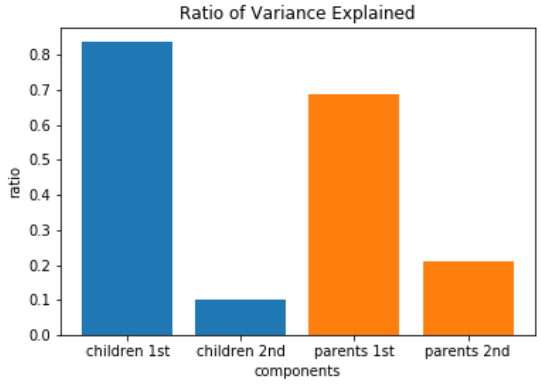
We built the matrix for the children and the parents respectively:

$$C = \left(Wage_{earnings} \;\; OtherLabor_{inc} \;\; SelfOrganizedBusiness_{inc} \;\; Property_{inc} \;\; Transfer_{inc}\right)$$

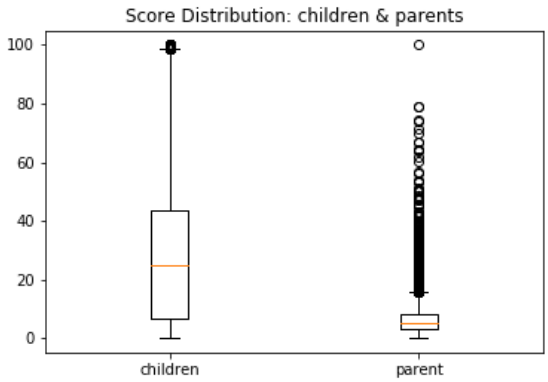
Each income source is a column vector, with the length equal to the sample size. And here we report the scatterplot of children and parents on the first and second components:



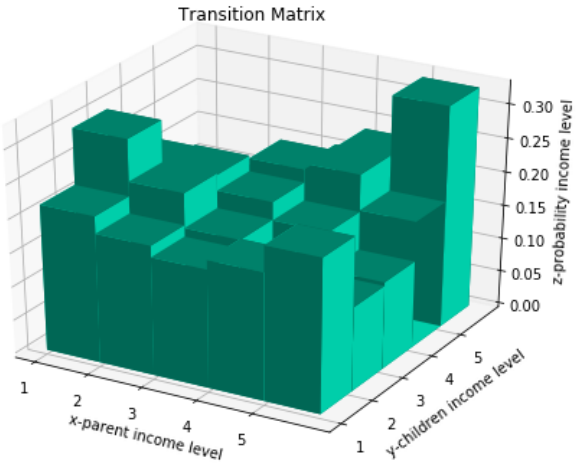




Our decomposition analysis was shown in these two plots. The dominating factor on the first component is the wage earnings, while the self-organized business income dominates the second component. For the parents, the first principle component is also dominated jointly by wage earnings and labor income from other sources, while the second component is essentially determined by transferred income. For the children, the first and second components were almost orthogonal, which implies that the wage earnings and SOB incomes are somewhat irrelevant. In contrast, the positive correlation between the two components of the parents is salient. In addition, from the 'Ratio of Variance Explained' plot, the dominance of the two principal components is overwhelming for the children. Therefore, we decide to blend the first and second component together, by taking the weighted average of the score on both components. The weights are their explaining ratio to the variance. Specifically, to mitigate the influence of outliers, which is quite salient from PCA plots above, we've decided to wise the part of data where children's economic scores fall below the 5 percentile or go beyond the 95 percentile.



Here we present the boxplots of the economic scores of parents and children. Children's scores are distributed more sparsely and evenly but have few outliers. However, the majority of the parents are at a relatively lower level of economic scores, and there is a huge number of outliers on the higher end.



Following the method from John et al (2004), then we divide up both generations into five layers, which were represented by *level 1* to *level 5*. And here we present the transition matrix, where the x-axis stands for the parents' economic level, y-axis stands for children, and the z-axis represents the empirical probability of transition from parents' level to the children's level accordingly. A rude observation is that children whose parents' earn less are more likely to move up. While for the rich parents, children mostly inherit their economic position, but the possibility of going to the bottom is also significant.

Reference

John P. Formby, W.James Smith, Buhong Zheng, 2004. Mobility measurement, transition matrices and statistical inference, *Journal of Econometrics*,Volume 120, Issue 1, 2004, Pages 181-205