

The Use of Discrete Data in PCA with Applications to Socioeconomic Status

Stanislav Kolenikov*

Gustavo Angeles†

March 3, 2004

*Department of Statistics, UNC Chapel Hill

†Carolina Population Center, UNC Chapel Hill

Structure of the talk

- Basic ideas of the principal component analysis
- Substantive problem: socio-economic status
- Discrete data problems
- Simulation study

Background

Concepts used in this talk:

- Multivariate statistics: principal components
- Social science methods: confirmatory factor analysis
- Econometrics: ordinal dependent variable methods

Not in this talk:

- Nonlinear PCA (in the spirit of Gifi's book)

What I also learnt:

- Logistics of a large scale simulations at `statapps.unc.edu`
- \LaTeX packages and styles: `seminar`, `hyperref`, `bibTeX`

Principal component analysis

Most mathematical statisticians view at PCA (Stat 185, Anderson (2003), Johnstone (2001)):

$$S \sim W_p(\Sigma, n) \Rightarrow S = \sum_{j=1}^p \hat{\lambda}_j \hat{a}_j \hat{a}_j^T, \quad \hat{a}_j \text{'s orthonormal} \quad (1)$$

with the main theoretical results shown for the samples from the multivariate normal distribution (Wishart S), such as asymptotic normality:

$$\sqrt{n}(\hat{\lambda}_j - \lambda_j) \rightarrow N(0, 2\lambda_j)$$

asymptotic independence of $\hat{\lambda}$'s from each other and from \hat{a}_j 's.

PCA as we know it - 2

Alternatively (and equivalently), it can be shown that if $\mathbb{E} x = 0$, $\mathbb{V} x = \Sigma$, then

$$\begin{aligned} \max_{\|a\|=1} \mathbb{V}[a^T x] &= \lambda_1, & a_1 &= \arg \max_{\|a\|=1} \mathbb{V}[a^T x], \\ \max_{\|a\|=1, a \perp a_1} \mathbb{V}[a^T x] &= \lambda_2, & a_2 &= \arg \max_{\|a\|=1, a \perp a_1} \mathbb{V}[a^T x] \end{aligned}$$

etc., so the principal components are the linear combinations of the variables that have the maximum variance, conditional on the previous PCs.

Yet another characterization is that of minimizing the norm of the residual vector from the linear fit to the data.

PCA as we know it - 3

Most applied statisticians view at PCA: whatever the data is, compute

$$S = 1/n \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T$$
$$C = \text{corr}(S), \quad S = \sum_{j=1}^p \hat{\lambda}_j \hat{a}_j \hat{a}_j^T \quad (2)$$

with primary interest on the first few components. a_j 's are usually referred to as *factor loadings* (and the original variables x_j as *factors*). The reason for standardizing the variables / applying the spectral decomposition to the correlation rather than to the covariance matrix is the likely difference in the scales of variables.

PCA as we know it - 4

Most common use of PCA: reduce the dimension of the indicator space. Also may help reveal interesting structure in the data on the scatterplots of the first PCs.

Measure of goodness of fit: the proportion explained by the first k components:

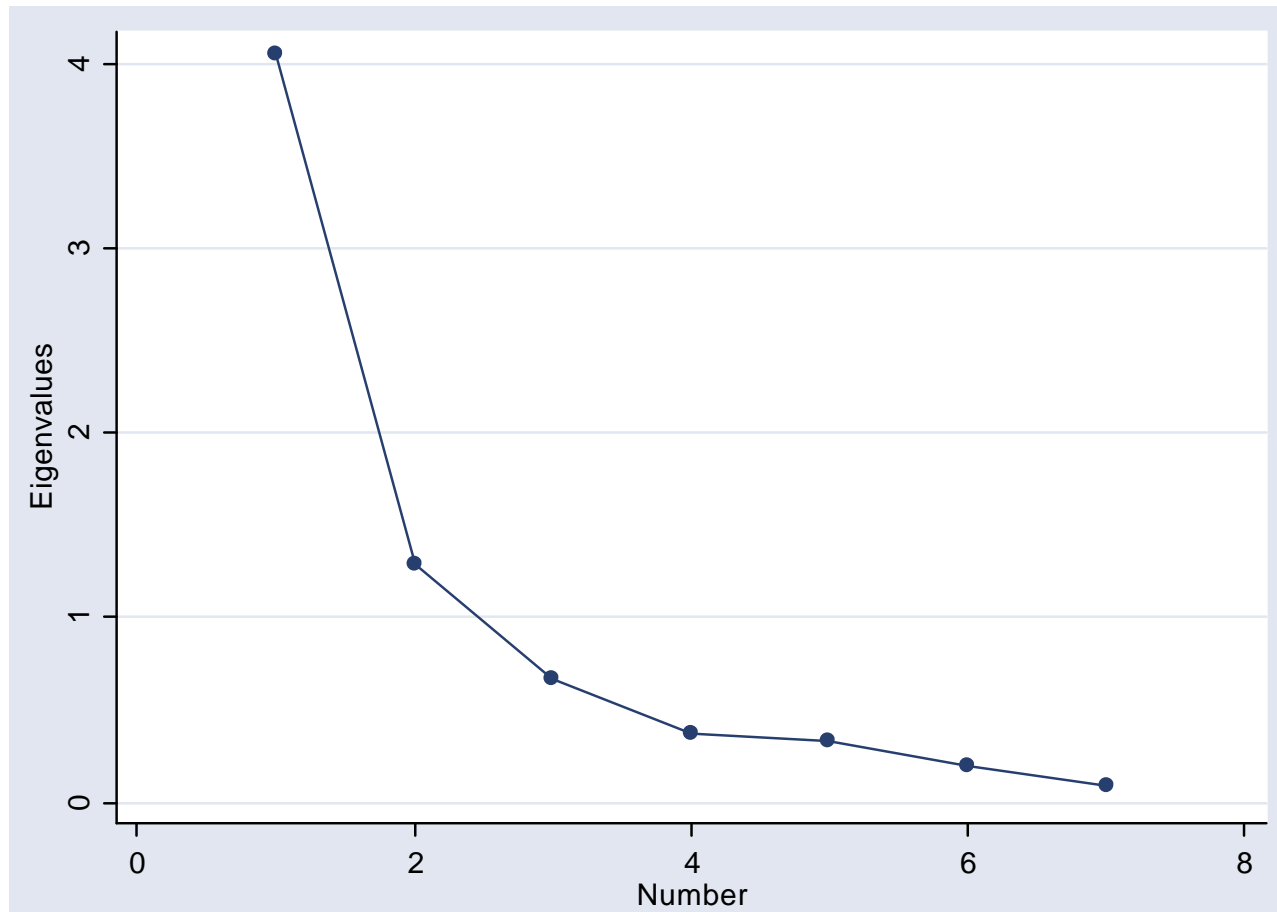
$$R_k = \frac{\lambda_1 + \dots + \lambda_k}{\lambda_1 + \dots + \lambda_p}$$

where λ 's may be both theoretical and empirical (i.e., $\hat{\lambda}_k$).

The graph of λ_k vs. k is usually referred to as the *scree* plot. It is helpful in eyeballing the number of “significant” components. (Otherwise, the number of “significant” components can be obtained from the distributional results through formal LR tests.)

PCA as we know it - 5

A typical scree plot



Factor analysis - 1

PCA is sometimes referred to as *(exploratory) factor analysis*. The aim of the factor analysis, in general, is to determine the primary factors underlying the variability of the observed data. Thus the interpretation of the PCA in terms of the factor analysis is that the first several “significant” principal components is that those factors orthogonal to each other provide the main sources of variability of the data, and the remaining components represent some sort of “white noise”.

Factor analysis - 2

There is a confirmatory version of factor analysis, too. One assumes that there is an underlying set of (possibly correlated) factors ξ that manifest themselves through some indicators

$$y = \Lambda\xi + \delta \tag{3}$$

where δ is the measurement error. Typically, all variables are assumed to be multivariate normal. Some sort of identification conditions need to be imposed on entries of Λ and $\Phi = \mathbb{V}\xi$ to assure unique maximizers of the likelihood exist, such as putting some of $\lambda_{ij} = 1$ or $\mathbb{V}\xi_k = 1$. Also, the model is usually considered in the deviations from the mean form, so that $\mathbb{E}\xi = 0$, $\mathbb{E}\delta = 0$.

Factor analysis - 3

Estimation:

$$\mathbb{V}[y] = \Lambda \mathbb{V}[\xi] \Lambda^T + \mathbb{V}[\delta]$$

Further identification conditions: # parameters $\leq p(p+1)/2$ where $p =$ # of indicators y .

If ξ , δ are assumed to be normal, then y 's are normal as well \Rightarrow maximum likelihood.

If not, asymptotically distribution free (minimum χ^2) methods can be used on individual entries of $\mathbb{V}[y]$.

Specification testing: form of LR test.

Socio-economic status

This is a multi-faceted concept that is supposed to capture many of the aspects of the relative position and achievements of an individual or a household in the society. It is believed to be determined by the resources available to the households (of which the primary is the pecuniary income), as well as the education levels attained by the members of the household, and (the prestige of) their occupations.

In turn, socio-economic status (SES for short) determines many individual and household decisions, including family planning, number of children, relocation decisions, etc.

The usual way to characterize SES is to construct an index which is simply a linear combination of some observed variables.

Related concepts: permanent income, household welfare.

SES - 2

Indicators of socio-economic status

- Household / individual income (equivalence scale?)
- Household / individual expenditure (equivalence scale?)
- Education of the household head / primary income earner
- Education of other members of the household
- Occupation of the household head / other members
- Property ownership: real estate, land, ...
- Quality of the residence
- Ownership of durable goods

SES - 3

Example: DHS (Demographic and Health Surveys), a standardized survey conducted in about 50 countries in the world. The SES related questions in Bangladesh (2000) study:

- Q18** What is the main source of water your household uses for dishwashing? (piped inside dwelling, piped outside dwelling, tube well, surface well, pond/tank/lake, river/stream, rainwater, other)
- Q19** What is the main source of drinking water for members of your household? (same categories)
- Q20** What kind of toilet facility does your household have? (septic/modern toilet, water sealed latrine, pit latrine, open latrine, hanging latrine, no facility/bush/field, other)

SES - 3a

Q22 Does your household (or any member of your household) have:

Electricity?	Wardrobe?	Table or chair?
Bench?	Watch or clock?	Cot or bed?
Working radio?	Working TV?	Bicycle?
Motorcycle?	Sewing machine?	Telephone?

Q24 Main material of the roof (recorded by the interviewer): natural roof, rudimentary (tin) roof, finished roof (cement, concrete, tiled), other

Q25 Main material of the walls (recorded by the interviewer): natural walls, rudimentary (wood) walls, brick/cement, tin, other

Q26 Main material of the floor (recorded by the interviewer): natural (earth) floor, rudimentary (wood) floor, finished (cement/concrete) floor, other

SES - 3b

Q27 Does your household own any homestead?

Q28 How much land does your household own (other than the homestead land)?

SES - 4

So if an index is to be constructed for SES, how does one specify the weights of the variables that enter the index? One can specify those normatively (such as a number of durable goods owned), or one can estimate the market values of those goods, or one can come up with entirely empirical way of getting the weights through factor analysis.

The indicators of SES can be used in (3) to estimate the socioeconomic status, at least in relative terms with other households / individuals in the sample.

Advantages of the CFA approach: allows specification testing.

Even simpler idea: run PCA on your data and take the first PC as the indicator of socioeconomic status / household welfare.

Advantages of the PCA: easier to compute, available in most statistical software.

SES - 5

Wait a second...

- What would income mean for an economy with mostly family farming?
- Didn't you say the y 's were assumed to be normal, to begin with?
- What about the indicator of possessing or not possessing a TV set or a refrigerator?
- Aren't income, expenditure, or acreage of land owned skewed?

The last point is relatively easy to take care of by means of transformations (take logs).

Discrete Data - 1

So, what can we do with discrete data? A few options

1. Ignore discreteness, use the moment correlations anyway
2. Create a binary variable for each category (why?). This is what the World Bank is doing (Filmer and Pritchett 1998)
3. Assume some sort of underlying distribution that was discretized into several categories — use *polychoric correlations* (Olsson 1979)

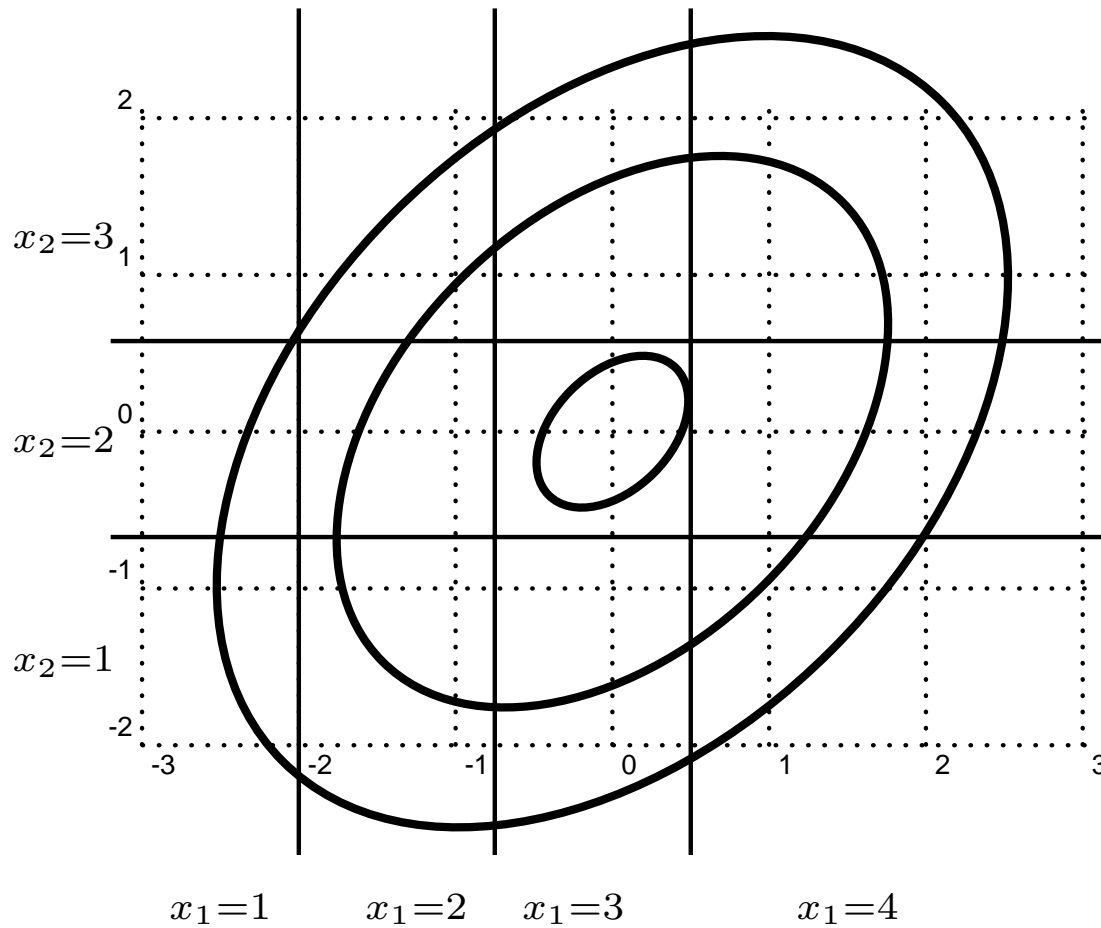
Is there any way to conceptualize how discrete (and, in particular, ordinal) variables appear in the analysis?

Discrete Data - 2

Suppose y_1^* , y_2^* are marginally standard normal with correlation ρ . They cannot be observed directly, but what is observable are their categorizations:

$$\begin{aligned}
 y_1 &= \begin{cases} 1, & \alpha_{1,0} = -\infty < y_1^* \leq \alpha_{1,1} \\ 2, & \alpha_{1,1} < y_1^* \leq \alpha_{1,2} \\ \vdots \\ k_1, & \alpha_{1,k_1-1} < y_1^* < \alpha_{1,k_1} = +\infty, \end{cases} \\
 y_2 &= \begin{cases} 1, & \alpha_{2,0} = -\infty < y_2^* \leq \alpha_{2,1} \\ 2, & \alpha_{2,1} < y_2^* \leq \alpha_{2,2} \\ \vdots \\ k_2, & \alpha_{2,k_2-1} < y_2^* < \alpha_{2,k_2} = +\infty \end{cases} \tag{4}
 \end{aligned}$$

Discrete Data - 3



Correlation: $\rho = 0.2$

Discrete Data - 4

How good is option 1 of dealing with the discrete data?

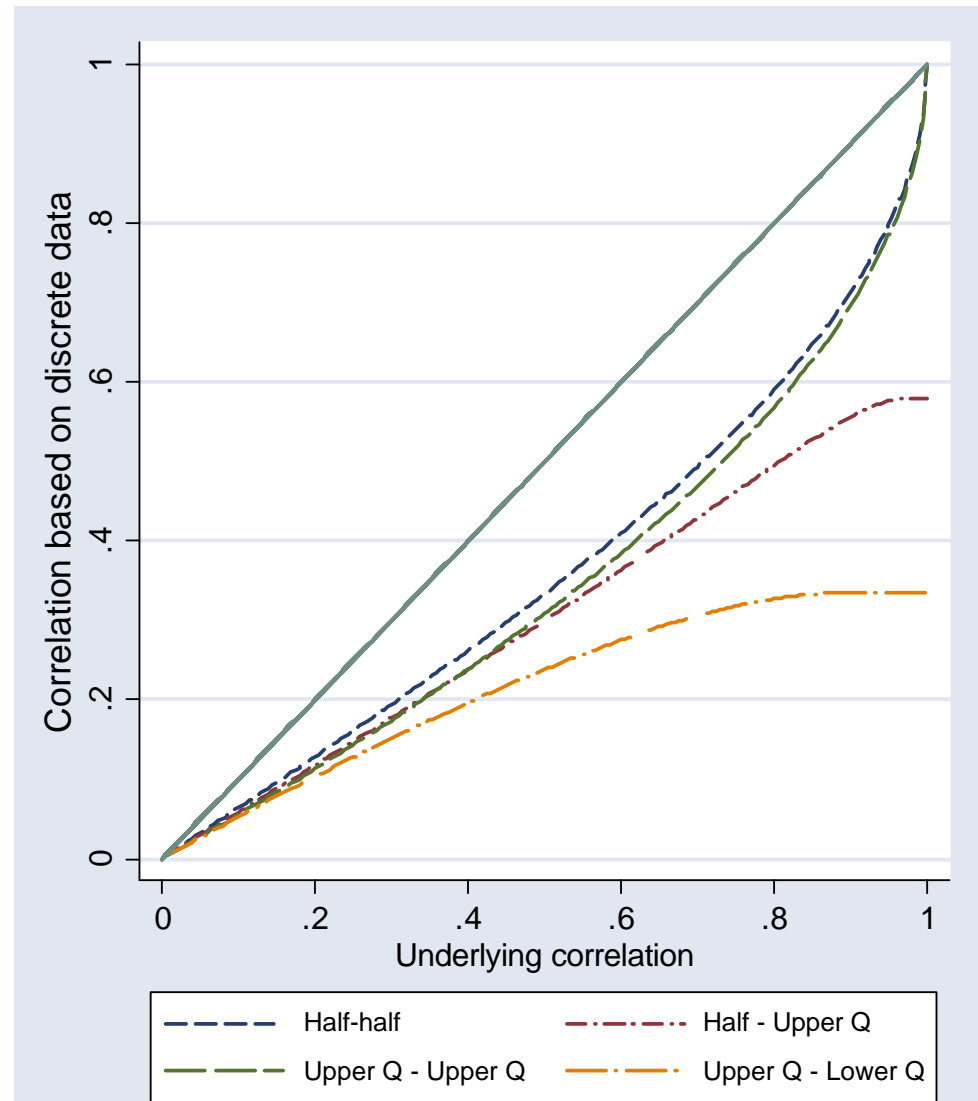
Let us compare the correlations of y_1^* , y_2^* with that of y_1 , y_2 , with just two categories. The four curves correspond to different combinations of α 's:

$$\alpha_{1,1} = \alpha_{2,1} = 0,$$

$$\alpha_{1,1} = \alpha_{2,1} = 0.66,$$

$$\alpha_{1,1} = 0, \alpha_{2,1} = 0.66,$$

$$\alpha_{1,1} = 0.66, \alpha_{2,1} = -0.66$$



Disrete Data - 5

What about option 2?

It can be argued that when all the data is broken into groups, and PCA is performed on the indicators of those groups, then the first PC connects the two largest groups, and the following components add other large groups in the order of their size. So if the two largest groups are the richest and the poorest, then the first PC roughly goes in the right direction, although the ordering of the households in between may be messed up.

Polychoric correlations - 1

The idea was first introduced by Pearson (1901), in the form of *tetrachoric* correlation for a 2x2 contingency table, and was then extended to multiple categories by Olsson (1979).

Denoting

$$\Phi_2(x, y; \rho) = \int_{-\infty}^x \int_{-\infty}^y \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}(u^2 - 2\rho uv + v^2)\right] du dv \quad (5)$$

the cdf of the standard bivariate distribution with the correlation of ρ , the probability of a cell in model (4) is

$$\begin{aligned} \pi(i, j; \rho, \alpha) &= \text{Prob}[y_1 = i, y_2 = j] = \\ &= \Phi_2(\alpha_{1,i}, \alpha_{2,j}; \rho) - \Phi_2(\alpha_{1,i-1}, \alpha_{2,j}; \rho) - \\ &\quad - \Phi_2(\alpha_{1,i}, \alpha_{2,j-1}; \rho) + \Phi_2(\alpha_{1,i-1}, \alpha_{2,j-1}; \rho) \end{aligned} \quad (6)$$

Polychoric correlations - 2

If the observations are independent, then the log likelihood can be written down as

$$\ln L = \sum_{i=1}^N \ln \pi(y_{i,1}, y_{i,2}; \rho, \alpha) \quad (7)$$

which can be maximized over ρ and α 's. The resulting ρ is what is referred to as the *polychoric correlation*.

In practice, the estimation most frequently is performed in two stages: first, the thresholds are estimated as

$$\alpha_{i,j} = \Phi^{-1} \left(\frac{-1/2 + \#\{y_i \leq j\}}{N} \right), \quad j = 1, \dots, k_i, \quad (8)$$

and then the correlation coefficient is estimated at the second stage by maximizing (7) conditional on α . This procedure does not yield the asymptotically normal estimates, but the discrepancies are really tiny and impractical (Maydeu-Olivares 2001).

Polychoric correlations - 3

In the multivariate setting (more than one ordinal variable), the full information maximum likelihood approach would be to estimate all correlation coefficients and all thresholds simultaneously. This is quite a tedious task (and quite computationally demanding, too), so the practical estimation procedure is (i) to compute the thresholds, as mentioned above, and then (ii) to estimate the correlation coefficients for each pair of variables and pool them together into a matrix.

(Open question: does this guarantee to produce a positive semidefinite matrix?)

Polychoric correlations - 4

How robust is the estimate of ρ to the departures from normality?

Formal check:

- Pearson χ^2 test on the cell proportions:

if $n_{i,j} = \# \text{obs: } \{y_1 = i, y_2 = j\}$, then

$$\sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \frac{(n_{i,j} - N\pi(i, j, \hat{\rho}, \hat{\alpha}))^2}{N\pi(i, j, \hat{\rho}, \hat{\alpha})} \xrightarrow{d} \chi_d^2$$

- Likelihood ratio test:

$$-2 \left(\sum_{i=1}^{k_1} \sum_{j=1}^{k_2} n_{i,j} \left[\ln \pi(i, j, \hat{\rho}, \hat{\alpha}) - \ln \frac{n_{i,j}}{N} \right] \right) \xrightarrow{d} \chi_d^2$$

The degrees of freedom for either of the tests is $d = k_1 k_2 - k_1 - k_2$.

Polychoric correlation - 5

In social statistics practice, those tests rarely reject the null hypothesis of normality. Besides, the polychoric correlation is not very sensitive for the violations of the distributional assumptions if the overall distribution is still elliptical (Kukuk 1998).

Polyserial correlation

If one of the variables (y_1) is ordinal and the other (y_2) is continuous, then the maximum likelihood estimate of the correlation between the two is referred to as the *polyserial*. It is the maximizer of the following log-likelihood (assuming that $y_2 \sim N(0, 1)$):

$$\ln L(\rho, \alpha; y_1, y_2) = \sum_{i=1}^N \sum_{j=1}^{k_1} \mathbf{I}[y_1 = j] (\Phi(\alpha_{1,j} - \rho x_2) - \Phi(\alpha_{1,j-1} - \rho x_2)) \phi(x_2) \quad (9)$$

PCA and SES

So the following PCA-based procedure for estimating the SES index can be suggested.

1. Estimate the MLE of the pairwise correlation between each of the two variables used in constructing the index. If the two variables are continuous, then this is the Pearson moment correlation. If the two variables are ordinal, then it this is the polychoric correlation. If one is ordinal and the other is continuous, then it is the polyserial correlation.
2. Run PCA on the resulting matrix, and interpret the first PC as the index of the household welfare.

How much of the performance do we buy by using the polychoric correlations rather than other methods?

Simulation - 1

To answer that question, a sizeable simulation project was entertained.

We generated the data according to the CFA model (3) with a single underlying factor in the regimes that were similar enough to the real world situations, and performed different analyses on the resulting data.

The simulation was performed on the server of statistical applications at UNC (<http://www.unc.edu/atn/statistical/>, the domain within a Sun E15K server with 20 processors and 40 GB of memory) as well as on several personal computers that the authors had access to. The software platform is Stata Special Edition, version 8.1. The project was spread into 28 separate threads. On average, a thread took about 1 to 3 days on a Pentium IV 1 GHz 256Mb RAM PC (single task), or 5 to 10 days on the multitask server. The likelihood maximization took the most of the simulation time.

Simulation - 2

The parameters and the settings of the simulation were as follows.

- Total number of indicators: from 1 to 10.
- The fraction of discrete variables: from 40% to 100%.
- The distribution of the underlying factor: normal; uniform; lognormal; mixture of two normals.
- The proportion of the variance explained: 80%; 50% if the total number of variables is greater than 5; 30% if the total number of variables is greater than 8.
- The values of λ : all ones; one of the discrete variables has $\lambda = 3$; one discrete and one continuous variables have $\lambda = 3$. This leads to the ratio in factor loadings of about 1.05–1.15, so the differences are not nearly as big as differences in λ 's.

- The number of categories of the discrete variables: from 2 to 8, with some gaps.
- The threshold settings: uniform (so that each group has the same number of observations); half observations are in the bottom category (heavy skewness and kurtosis, at least for a large number of categories), or 60/40 for two categories; half observations are in the central category (high kurtosis with low skewness); random thresholds.
- The sample sizes: 100, 500, 2000, 10000.
- The analyses performed: PCA on the ordinal categorical variables; PCA on the dummy variables corresponding to the categories; PCA on the ordinal variables with the number of the category replaced by the group means; PCA of the polychoric correlation matrix; PCA on the original y^* variables as the benchmark.

For each combination of settings, only one Monte Carlo sample was taken.

Simulation - 3

The total number of observation: 969395, with 49 variables describing the settings and the outcomes (Stata file size is 298 Mbytes).

The primary outcome variables are the externally and internally defined goodness of fit measures, namely, the rank correlations of the first PC with the original factor ξ_1 , and the reported proportion of the explained variance. The motivation behind the focus on those quantities is that the rank correlations shows how good is this measure for ranking individuals (which is what the PCA is used for in the well-being analysis), and the proportion explained is taken to be the primary (and often the only) indication of how strongly the variables are correlated, and how much information is contained in the first PC.

Results - 1

Several “sections” of this data set:

- No continuous variables, greatest losses to discreteness
- Maximal number of variables, effects of continuous variables and proportion explained
- Binary data, greatest losses to categorization

Results - 2

Findings are qualitatively, and if not quantitatively, similar.

- Greater proportion of the variance explained by the underlying factor ξ_1 improves both actual and reported fit
- More variables improves fit, the improvements fading off for about 6 or 8 variables when the explanatory power is high ($R_1 = 0.8$), but still there for 10 variables when it is lower
- More categories is usually good for polychoric correlations, but bad for PCA on category indicators
- Fit is substantially worse for the lognormal distribution of the factor than for any other setting (higher kurtosis?)
- Only the polychoric correlation based methods are able to consistently estimate the proportion of explained variance

See also

More details can be found in Kolenikov and Angeles (2004), available upon request.

References

- Anderson, T. W. (2003) *An Introduction to Multivariate Statistical Analysis*, 3rd edition ed. (New York: John Wiley and Sons)
- Filmer, Deon, and Lant Pritchett (1998) ‘Estimating wealth effect without expenditure data — or tears: An application to educational enrollments in states of india.’ World Bank Policy Research Working Paper No. 1994, The World Bank, Washington, DC
- Johnstone, Iain M. (2001) ‘On the distribution of the largest eigenvalue in principal component analysis.’ *Annals of Statistics* 29, 295–327
- Kolenikov, Stanislav, and Gustavo Angeles (2004) ‘The use of discrete data in PCA: Theory, simulations, and applications to socioeconomic indices.’ Technical Report, MEASURE/Evaluation project, Carolina Population Center, University of North Carolina, Chapel Hill
- Kukuk, Martin (1998) ‘Analyzing ordered categorical data derived from elliptically symmetric distributions.’ Technical Report, University of Tübingen,

Germany

- Maydeu-Olivares, Albert (2001) 'Testing categorized bivariate normality with two-stage polychoric correlation estimates.' Technical Report, University of Barcelona, Dept. of Psychology
- Olsson, U. (1979) 'Maximum likelihood estimation of the polychoric correlation.' *Psychometrika* 44, 443–460
- Pearson, Karl (1901) 'Mathematical contributions to the theory of evolution. vii. on the correlation of characters not qualitatively measurable.' *Philosophical Transactions of the Royal Society of London, Series A* 195, 1–47