The Use of Discrete Data in Principal Component Analysis for Socio-Economic Status Evaluation

Stanislav Kolenikov[a] **Gustavo Angeles**[b]

February 2, 2005

[a] Department of Statistics, UNC Chapel Hill, and Centre for Economic and Financial Research, Moscow

[b] Department of Maternal and Child Health and Carolina Population Center, UNC Chapel Hill

Outline

1. Motivation for socio-economic status (slide 3)

   Who is interested in SES, and why?

2. Principal component analysis (slide 11)

   Is this a reasonable procedure to generate weights for SES index?

3. Applications: Bangladesh DHS+, 2000 (slide 23) and Russia, RLMS 1994–2001 (slide 34)

   Does it work for developing countries? Does it work for middle income countries?

   Does it work with binary data only?

4. Monte Carlo study of the different flavors of PCA (slide 40)

   Can we make any general conclusions about the methods?

5. Conclusions and references (slide 48)

   How much room is there for improvement?

Motivation: Socio-Economic Status

- Used to identify groups of people who share a similar position with relation to the involvement in social networks and access to economic resources
- SES is of interest for health economists
  - Household decision making (fertility, education, relocation . . .) and outcome (mortality, service use, . . .) variable
  - Economic policy variable (project allocation)
- What's in there?
  - Income and wealth
  - Education
  - Occupation and its prestige
  - Varies by culture

SES measurement

SES is a multifaced concept, no direct measure available

- Income: good, but not the only aspect; data not always of appropriate quality; varies a lot; saving and borrowing?
- Consumption/expenditure: less variable than income; otherwise may have the same problems; use of durable goods? out-of-market transactions?
- Single proxy: unreliable?
- Aggregation of several indicators: weights?

$$SES_i = \sum_k w_k x_{ik} \tag{1}$$

See Bollen, Glanville & Stecklov (2001), Bollen, Glanville & Stecklov (2002*a*), Bollen, Glanville & Stecklov (2002*b*).

On top of everything, the issues of endogeneity (Thomas & Strauss 1995).

SES measurement - 2

The common ways to aggregate several indicators into a single 1D measure (or, in other words, to arrive at weights w_k in (1)):

- w_k = the value of an asset (self-reported; median value; external estimate)
- $w_k = 1$ gives the sum of assets (why a car should have the same weights as a radio?)
- w_k are determined by PCA
  - Filmer & Pritchett (2001): break every categorical variable into dummies
  - Kolenikov & Angeles (2004): use ordinal variables, maybe through the polychoric correlations

SES measurement - 3

Demographic and Health Survey, Bangladesh, 2000.

9753 observations in 341 clusters. The SES household level variables are:

`hv201`	source of drinking water
`hv202`	source of non-drinking water
`hv205`	type of toilet facility
`hv206`	has electricity
`hv207`	has radio
`hv208`	has television
`hv210`	has bicycle
`hv211`	has motorcycle
`hv213`	main floor material
`hv214`	main wall material
`hv215`	main roof material

Website with data: http://www.measuredhs.com

SES measurement - 4

Russian Longitudinal Monitoring Survey (38 clusters, $\sim$ 3600 households)

- consumption, income
  - any single round
  - aggregate over several rounds
- assets
  - refrigerator
  - freezer
  - washer
  - black & white TV
  - color TV
  - VCR
  - computer
  - car
  - truck
  - motorcycle or boat
  - tractor
  - dacha
  - more than one apartment
  - floor and living space (m^2)
  - central heating
  - central water
  - hot water
  - metered gas
  - telephone
  - central sewerage

Linear combinations - 1

Why do people love linear representations like (1)?

Under normality (Mardia, Kent & Bibby 1980),

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_Y \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \Sigma_{YX} \\ \Sigma_{YX}^T & \Sigma_{XX} \end{pmatrix} \right) \Longrightarrow$$

$$\Longrightarrow Y|X \sim N\Big(\mu_Y - \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X), \sigma_Y^2 - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{YX}\Big) \tag{2}$$

so that

$$\mathbb{E}[Y|X] = \mu_Y - \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X)$$

is a linear combination of the components of X vector.

Linear combinations - 2

One special case of (2): linear regression

$$Y = X^T \beta + \epsilon$$

Nothing is said about Σ_{XX}. The randomness is in ϵ, the regression error. Σ_{YX} and σ_Y^2 are derived from β's.

Linear combinations - 3

Another special case of (2): factor models

$$X_k = \lambda_k Y + \delta_k \tag{3}$$

where Y is unobserved factor, λ_k are factor loadings, and δ_k are measurement errors.

Note:

- X's are dependent variables (indicators)
- The covariance structure Σ_{YX}, Σ_{XX} is derived from the model (3)
- Prediction $\hat{Y}_i = \text{const} + \sum_k w_k X_{ik}$ is linear in X_i

Principal components - I

One of the historically oldest ways to aggregate several indicators into a single measure is the use of *principal components*. The principal components of variables $x_1, \ldots, x_p$ are linear combinations $\mathbf{a}_1'\mathbf{x}, \ldots, \mathbf{a}_p'\mathbf{x}$ such that

$$\mathbf{a}_1 = \arg \max_{\mathbf{a}:\|\mathbf{a}\|=1} \mathbb{V}\left[\mathbf{a}'\mathbf{x}\right],$$

$$\vdots$$

$$\mathbf{a}_k = \arg \max_{\substack{\mathbf{a}:\|\mathbf{a}\|=1, \\ \mathbf{a}\perp\mathbf{a}_1,\ldots,\mathbf{a}_{k-1}}} \mathbb{V}\left[\mathbf{a}'\mathbf{x}\right] \tag{4}$$

Principal components - II

Solution is found through the eigenproblem for $\Sigma = \mathrm{Cov}[\mathbf{x}]$:

$$\text{find } \lambda, \mathbf{v} \neq 0 \text{ s.t.} \tag{5}$$

$$\Sigma\mathbf{v} = \lambda\mathbf{v} \tag{6}$$

Certain linear algebra properties: Σ is p.s.d. $\Longrightarrow \lambda > 0$; uniqueness; orthogonality of eigenvectors

Certain asymptotic results for sample covariance matrices: asymptotic normality

Principal components - III

Properties and features

++ Standard multivariate statistical analysis technique; taught in most multivariate statistics classes, some econometric classes, some quantitative social sciences classes

++ Available in most statistical packages

– – Developed and suitable for continuous (ideally, mutlivariate normal) data

– – Still a black box in applied research and policy advice?

References: Pearson (1901*b*), Hotelling (1933), Mardia et al. (1980), Rencher (2002)

Principal components - IV

** Principal component analysis works on the covariance or correlation matrix to extract the directions in the multivariate space that is the "most informative", which means, have the greatest variability.

++ Usually, a few first components explain most of the variability in the data

** Mathematics of PCA: eigenvalue problem

– – Does not take into account non-normalities of the data, such as nontrivial skewness, kurtosis, or discreteness

Discrete data

Types of discrete data

- count data (# of children, # of rooms, # of accidents in a month)
- nominal data (gender, industry, occupation, employment status)
- ordinal data (Likert scales for degree of agreement, education, quality of house materials, ownership of a good)

PCA with discrete ordinal data

- ordinal PCA: ignore discreteness
- Filmer-Pritchett procedure: break down categories into dummy variables
- polychoric PCA: use the polychoric correlation matrix
- group means: use means of a truncated distribution for variable scores

Ordinal PCA

Suggestion: perform PCA on the original variables, completely ignoring their discreteness

+ Very easy to do, although may need recoding of the data to a Likert scale

– – Correlations are on a smaller side

– – Distributional assumptions for PCA are violated; high skewness and kurtosis $\Rightarrow$ different asymptotic properties

Filmer and Pritchett procedure

Filmer & Pritchett (2001) suggested generating dummy variables for each of the categories — most likely, following a common suggestion that a categorical variable should be treated that way when it is used in regression. (See however discussion on slide 10.)

– – The ordering of values of an ordinal variable is lost

– – Extra correlations are introduced into data: instead of concentrating on figuring out the relations between different measures of SES, the PCA now has to work on the correlations polluted by the negative relations between the variables produced from a single ordinal source variable

++ Imposes fewer assumptions on the data — allows to determine the "true"(?) ordering of categories

Polychoric correlation - 1

Suppose x_1^*, x_2^* are jointly bivariate normal with standard normal marginals and correlation ρ. Further, the ordinal x_1, x_2 are obtained by discretizing x_1^*, x_2^* according to the set of thresholds $\alpha_{k1}, \ldots, \alpha_{k,K_k-1}$:

$$x_k = r \text{ if } \alpha_{k,r-1} < x_k^* < \alpha_{k,r} \tag{7}$$

where $\alpha_{k,0} = -\infty$, $\alpha_{k,K_k} = +\infty$. Then if

$$\Phi_2(s,t;\rho) = \int_{-\infty}^{s}\int_{-\infty}^{t} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\Big[-\frac{1}{2(1-\rho^2)}\left(u^2 - 2\rho uv + v^2\right)\Big]\, du\, dv \tag{8}$$

is the cdf of the bivariate standard normal distribution, then the cell probability is

$$\begin{aligned}\pi(i,j;\rho,\alpha) &= \mathrm{Prob}[x_1 = i, x_2 = j] = \\ &= \Phi_2(\alpha_{1,i}, \alpha_{2,j};\rho) - \Phi_2(\alpha_{1,i-1}, \alpha_{2,j};\rho) - \\ &-\Phi_2(\alpha_{1,i}, \alpha_{2,j-1};\rho) + \Phi_2(\alpha_{1,i-1}, \alpha_{2,j-1};\rho)\end{aligned} \tag{9}$$

Polychoric correlation - 2

The maximum likelihood estimate of ρ can be obtained from discrete data by maximizing

$$\log L(\rho, \alpha; X) = \sum_{i=1}^{n} \log \pi(x_{i1}, x_{i2}; \rho, \alpha)$$

In practice, the estimates are obtained from bivariate information maximum likelihood procedure (BIML):

1. estimate $\hat{\alpha}_{k,\cdot}$ from the marginal distribution of x_k;
2. estimate $\hat{\rho}_{kl}$ conditional on those thresholds;
3. populate the correlation matrix $\mathrm{Corr}[X]$;
4. perform further analysis on this polychoric correlation matrix (e.g. PCA)

Polychoric correlation - 3

→ Pearson (1901*a*), Olsson (1979), Jöreskog (2004)

** Involves two ordinal variables

** Assumes an underlying bivariate normal distribution with cutoff points, similar to ordered probit regression

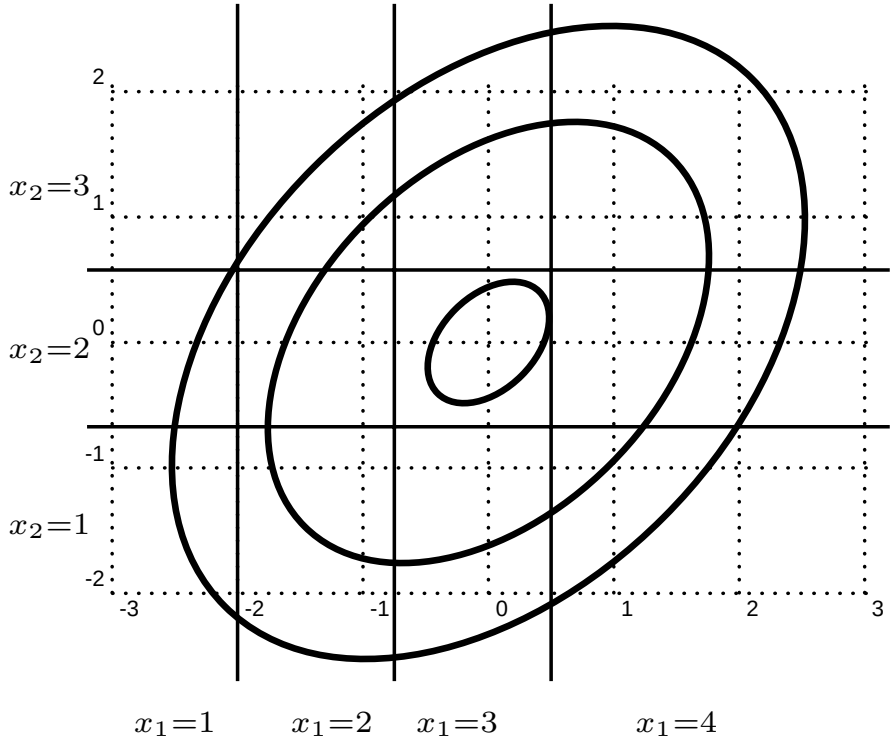
++ Is a maximum likelihood estimate of the correlation of that underlying bivariate normal distribution: asymptotically efficient

––– Requires iterative maximization, hence slow, especially in large data sets and with many variables. (Bangladesh: ∼10 minutes on a 1.5GHz PC!!!) May have convergence difficulties even with 100s of observations.

–– Estimation routines available only in specialized software.

++ `polychoric` Stata package developed in-house at CPC

Polychoric correlation - 4



$\alpha_{1,1} = -2$, $\alpha_{1,2} = -0.75$, $\alpha_{1,3} = 0.5$; $\alpha_{2,1} = -0.25$, $\alpha_{2,2} = 1$, and the correlation of the underlying bivariate normal is 0.2.

Group means method

** Idea: use $\mathbb{E}[x_k^*|x_k = j]$ as a score value for category j of variable x_k

– – Need some distributional assumptions, such as normality

++ Once computed, easy to use in the standard PCA routines

– – Totally *ad hoc*

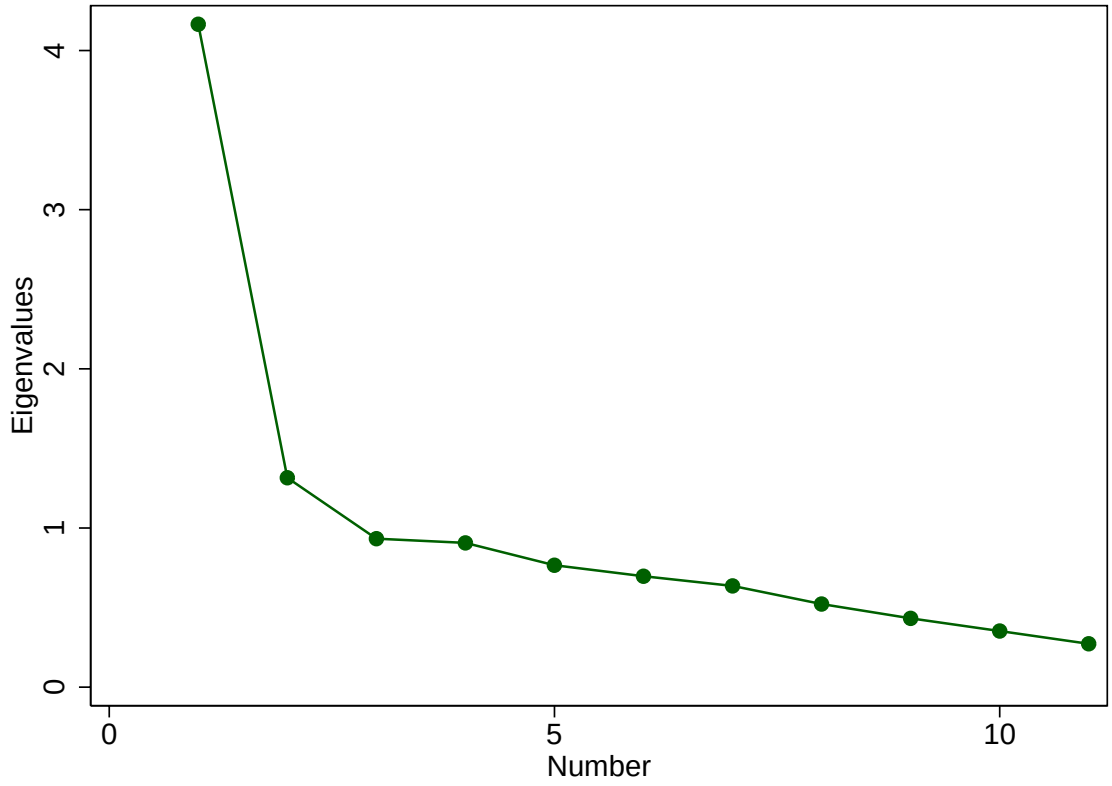
Bangladesh 2000 (†), ordinal PCA - 1

PCA is performed on recoded asset ownership variables, recoded to have the range of 1 to about 5 (Likert scale), “higher” meaning “better”.

Interpretations of the results:

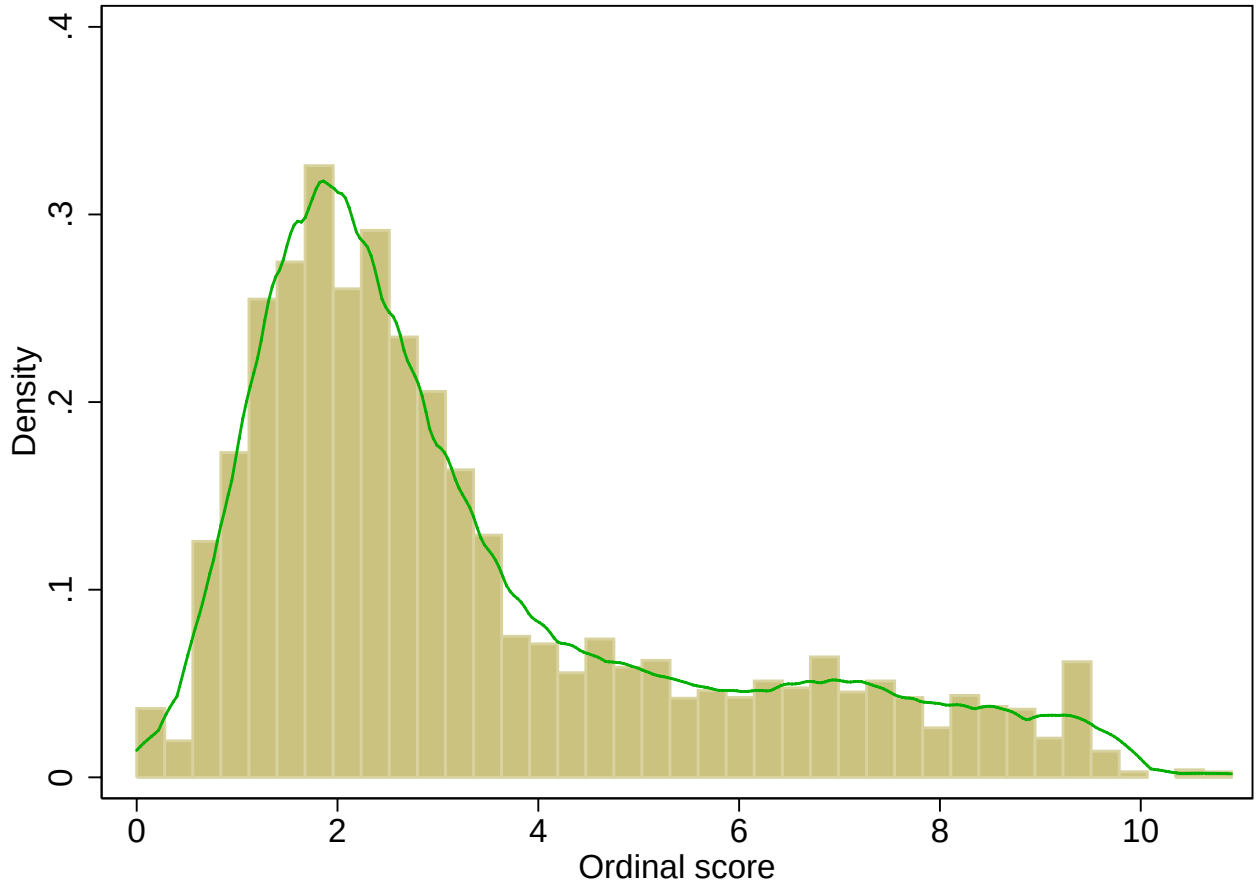
- Eigenvalues: 41% of the variability is reportedly explained by the first component
- Eigenvalues: other components are probably noise (see the graph)
- Loadings: all coefficients of the 1st PC are positive: having an asset increases SES
- Loadings: having a bike or a motorcycle is not as important as others
- PC score: skewed, as wealth distributions should be; Gini = 0.3665
- PC score: some lumping of the observations together; the three most populated categories account for about 10% of the data

Bangladesh 2000, ordinal PCA - 2



Scree plot: the first component is significant, the second component is marginally significant. The proportion of explained variance is 38%.

Bangladesh 2000, ordinal PCA - 3

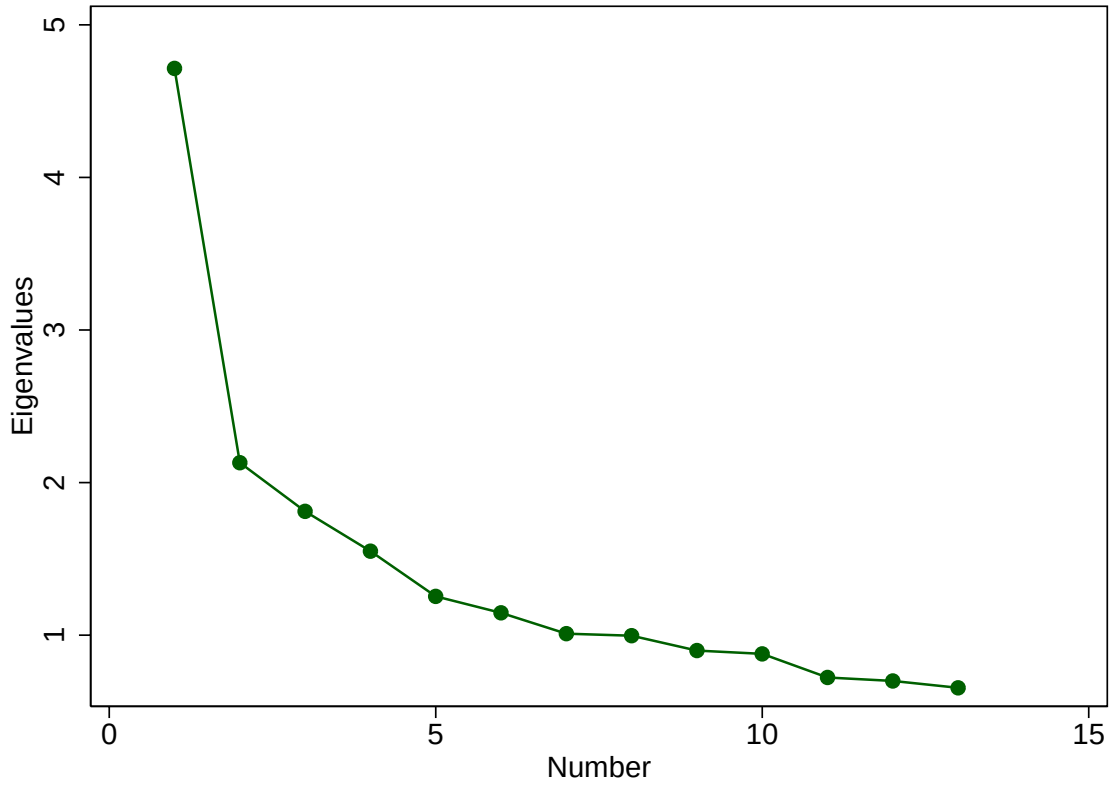


Distribution of the first PC score.

Bangladesh 2000: Filmer-Pritchett PCA - 1

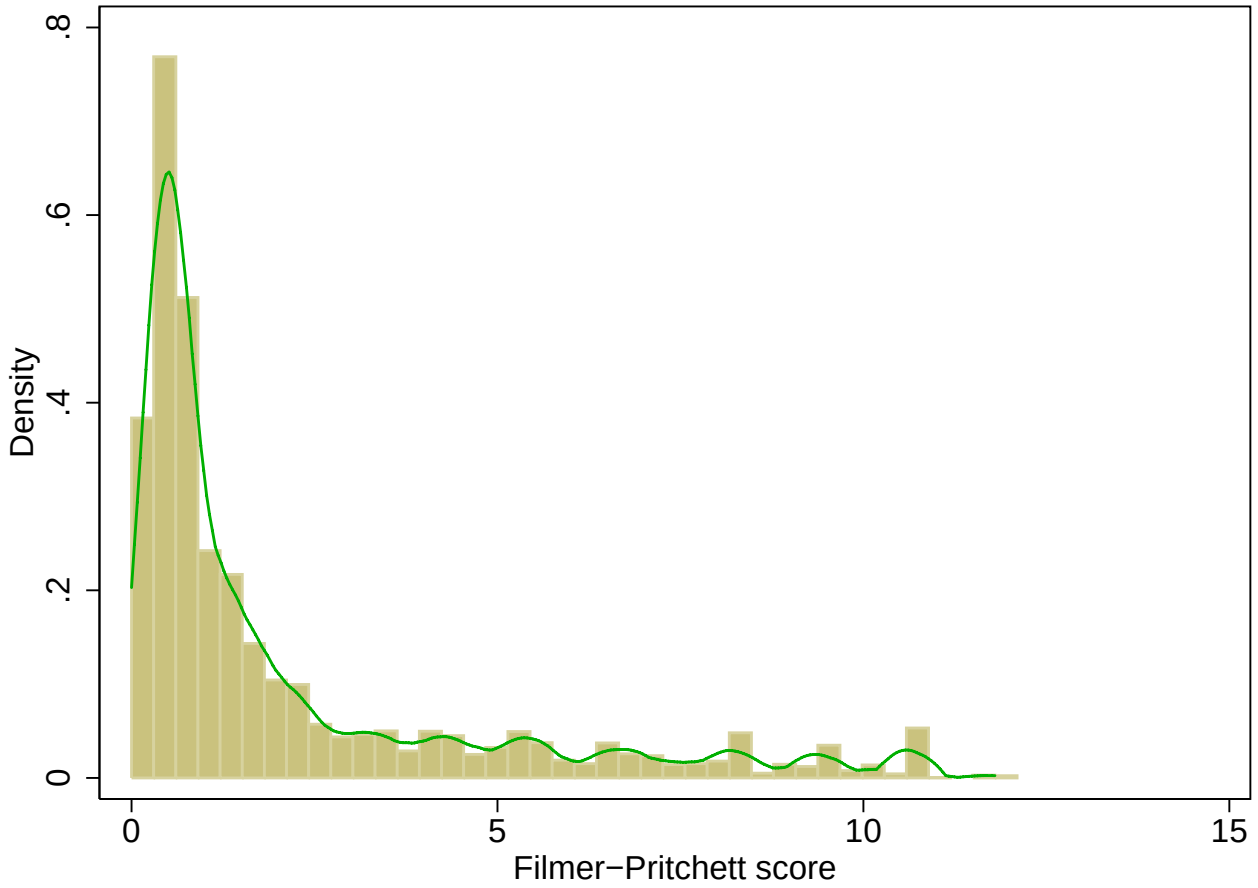
- Eigenvalues: 22% of the variability is reportedly explained by the first component
- Eigenvalues: high noise in other components; some five components are more informative than the remaining noise (derivatives of the ordinal variable?)
- Loadings: not positive in the first PC, although ordering is mostly monotone in concordance with expectations
- Loadings: no direct comparisons of relative importance is possible
- PC score: lumping is the same
- PC score: skewed, Gini = 0.58.
- PC score: multimodality in the upper end of distribution

Bangladesh 2000, Filmer-Pritchett PCA - 2



Scree plot: the first component is significant, the next three are probably significant, too. The proportion of explained variance is 22%.

Bangladesh 2000, Filmer-Pritchett PCA - 3



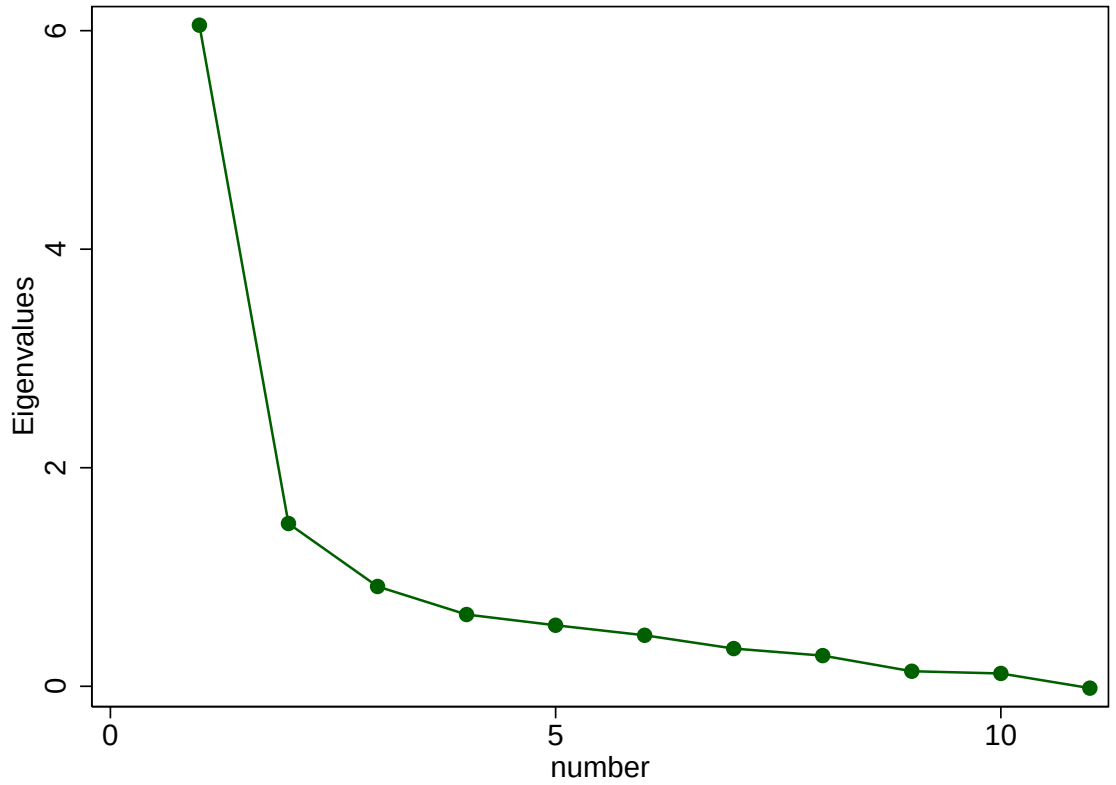
Distribution of the first PC score.

Bangladesh 2000: polychoric PCA - 1

- Eigenvalues: 55% of the variability is reportedly explained by the first component
- Eigenvalues: other components are probably noise
- Loadings: monotone pattern of the category scores by design
- PC score: skewed, as wealth distributions should be; Gini = 0.33
- PC score: same lumping

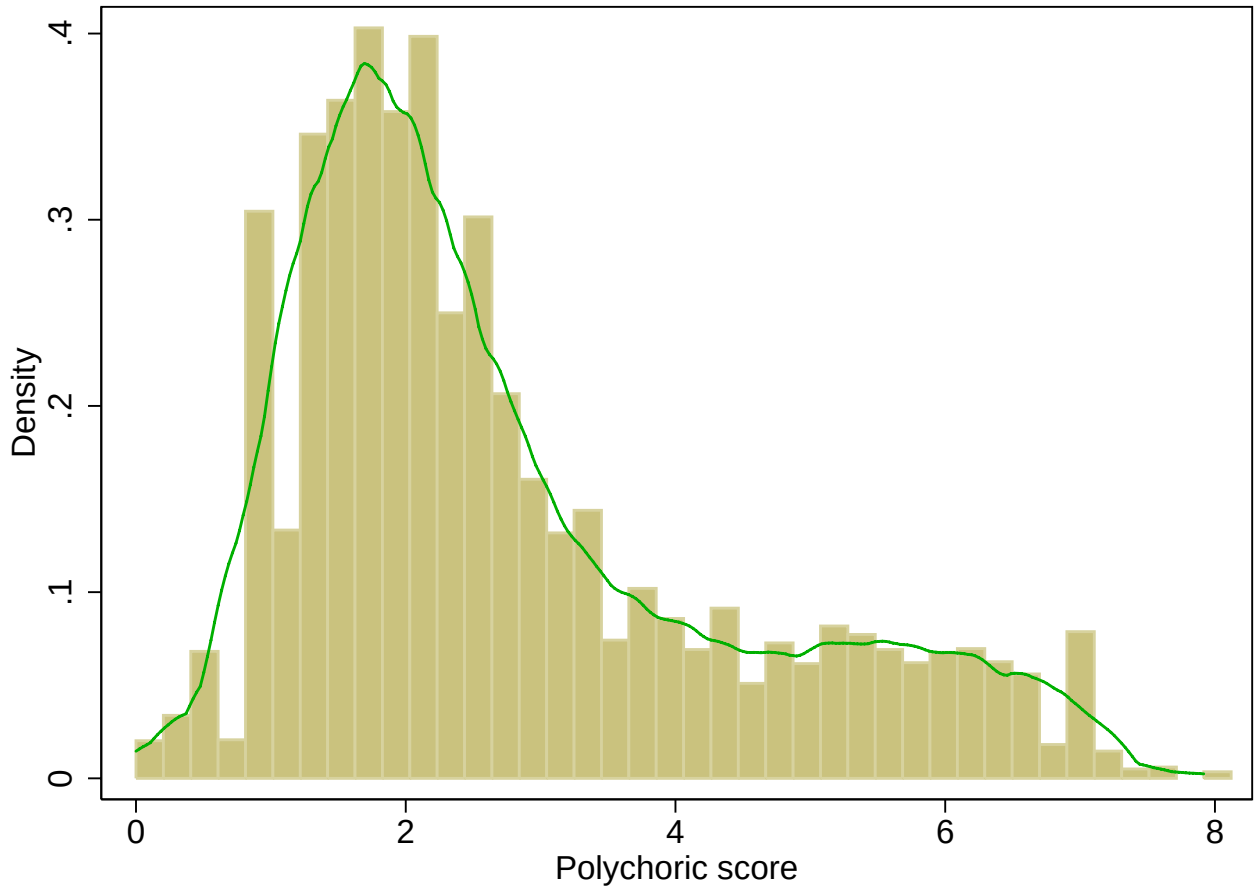
Overall, the results are quite similar to the PCA on ordinal variables.

Bangladesh 2000, polychoric PCA - 2



Scree plot: the first component is significant, the others are probably noise. The proportion of explained variance is 55%.

Bangladesh 2000, polychoric PCA - 3



Distribution of the first PC score.

Bangladesh 2000: Comparison - 1

- All three procedures produced 1336 unique values in 9753 observations; the largest lump of identical scores (due to identical values of indicators) has 383 observations.
- Rankings are very similar for the polychoric and ordinal methods, but the quintiles are mixed between any of those two and the Filmer-Pritchett method
- Kendall's τ between the polychoric and the F-P scores is 0.54, which means that about 23% of the pairs of observations are discordant (one of the observations is scored higher than the other by one method and lower by the other)
- The Filmer-Pritchett and polychoric/ordinal methods do not place *any* of the households jointly into the first quintile. The quintile cross-classifications are quite inconsistent

Bangladesh 2000: Comparison of the procedures - 2

Factor loadings: see Table 2 (page 13) of empirical paper.

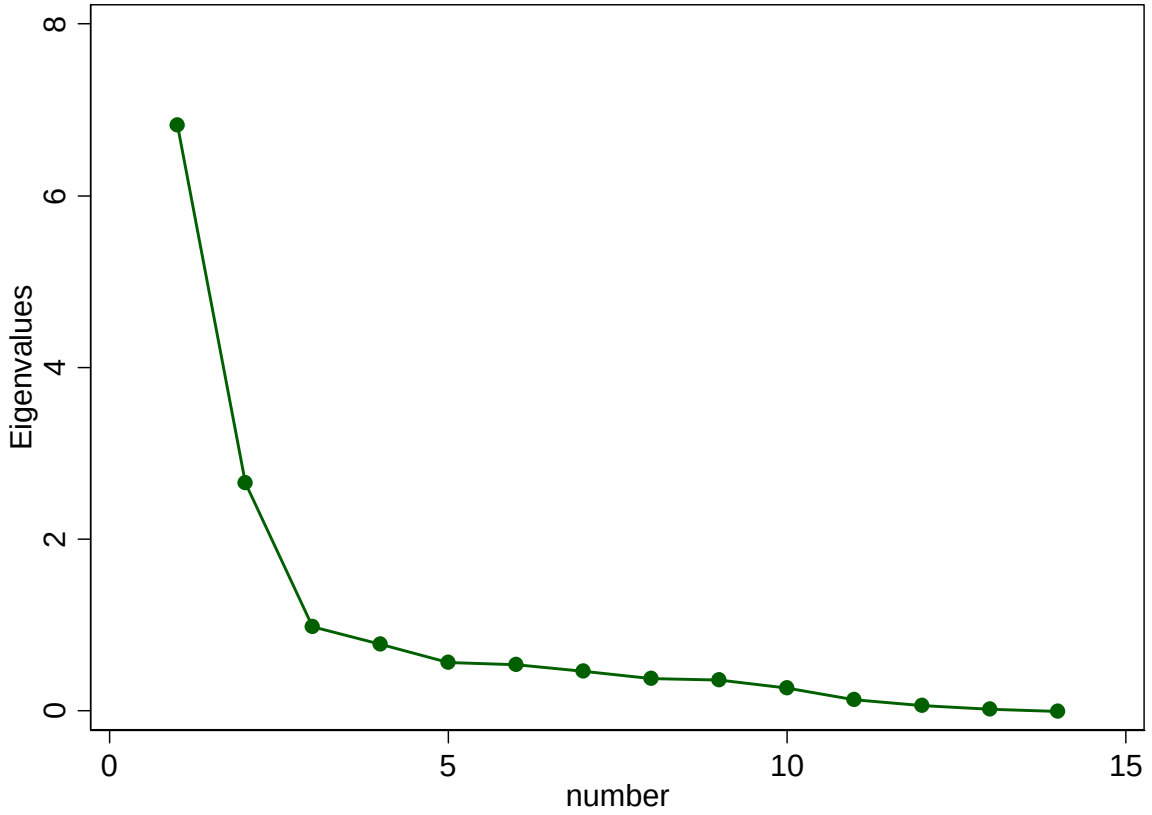
Scatter plot of scores: see Figure 4 (page 16) of empirical paper.

Russia 1994–2001, RLMS (†): polychoric PCA

PCA is performed on the ownership variables (excluding ownership of a truck, motorcycle or boat, tractor, 2nd apartment, black & white TV, and living space):

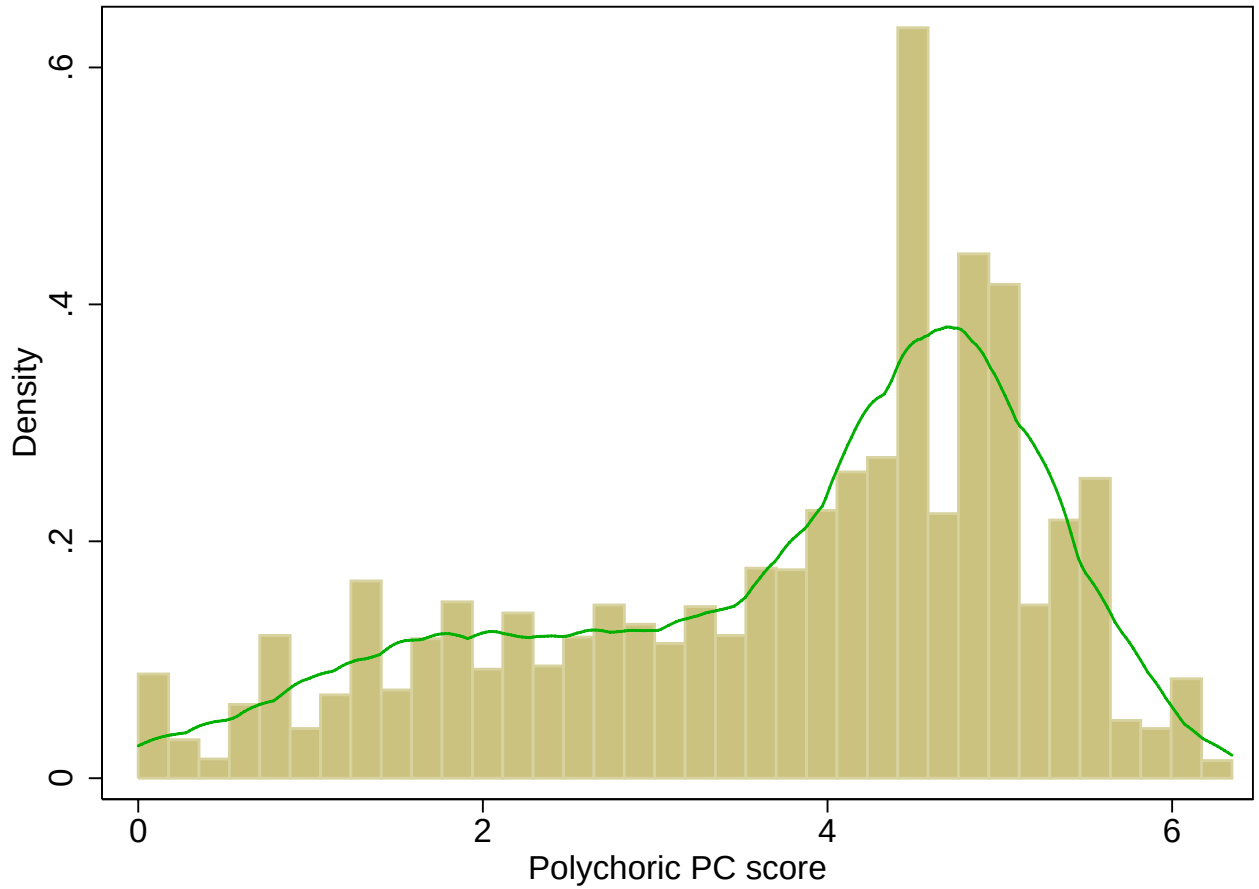
- proportion of explained variance = 48.7 %
- two components are significant; the second component highlights access to utilities, as well as ownership of some "trivial" items (79% of HH have a washer; only 27% of population have a car)
- loadings: utilities are most important, as well as owning a fridge, color TV, and computer
- PC score: *negative* skewness; Gini = 0.215
- PC score: some lumping, 584 unique values, 280 HHs in the largest one (6.7%)

RLMS 1994–2001: polychoric PCA - 2



Scree plot: the first two components are significant. Proportion of variance explained by the first component is 49%.

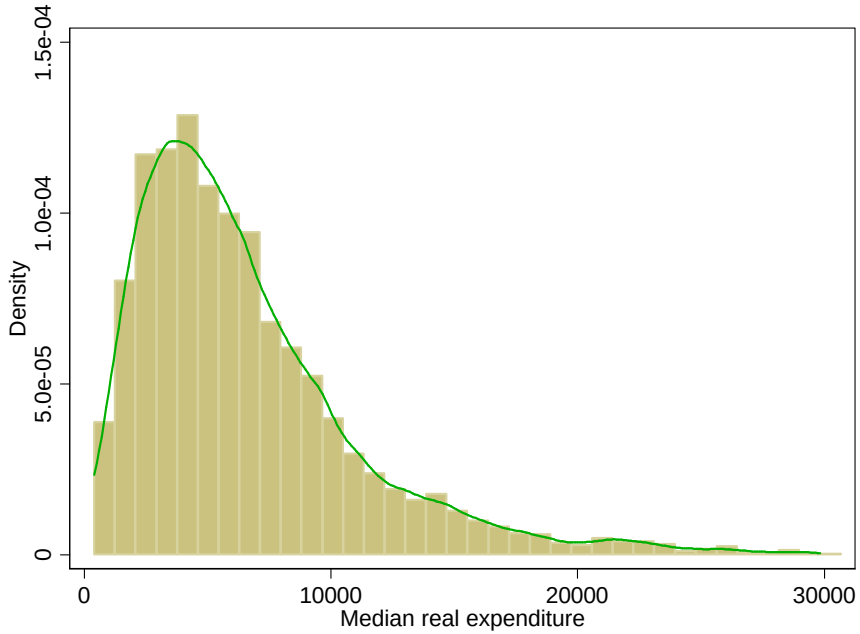
RLMS 1994–2001: polychoric PCA - 3

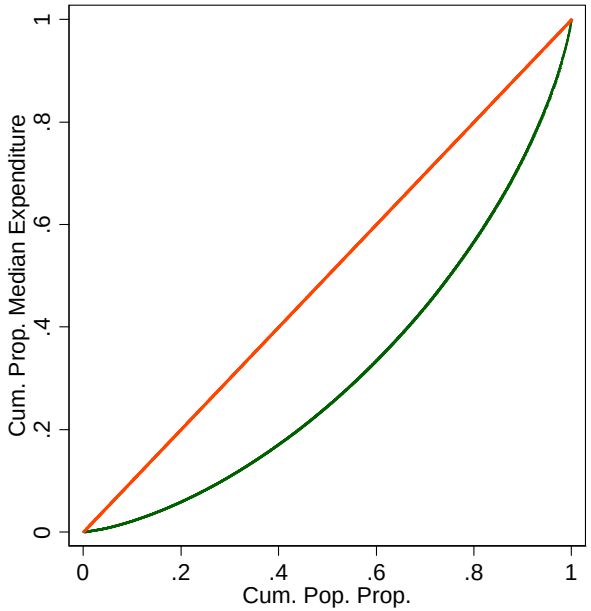


Distribution of the first PC score.

RLMS 1994-2002, median expenditure

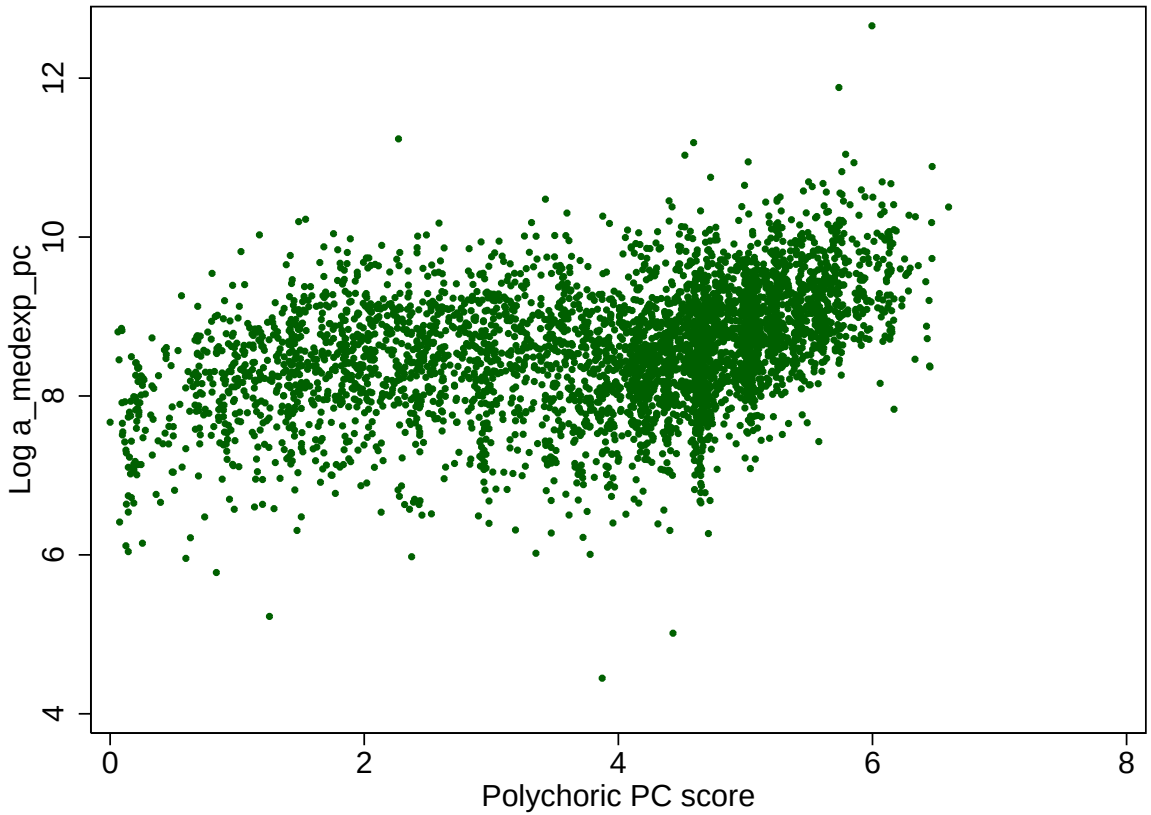
The panel character of the data allowed to compute the measure of permanent consumption — the median expenditure for the household that appear frequently enough, in at least three waves (4190 HHs).





Gini of median expenditure = 0.372.

RLMS 1994–2001, comparison - 1



Correlation of log median expenditure per capita with the polychoric score is 0.3510.

RLMS 1994–2001, comparison - 2

- Ordinal PCA $\equiv$ Filmer-Pritchett PCA
- Ordinal PCA scores $\approx$ polychoric PCA scores (Kendall's $\tau = 0.96$)
- Unsatisfactory performance against "permanent consumption"
  - Many acquired under Soviet regime, and not indicative of the current SES?
  - Items irrelevant for top SES?

Monte Carlo study

Data generating model: confirmatory factor analysis (3) with discretization (7).

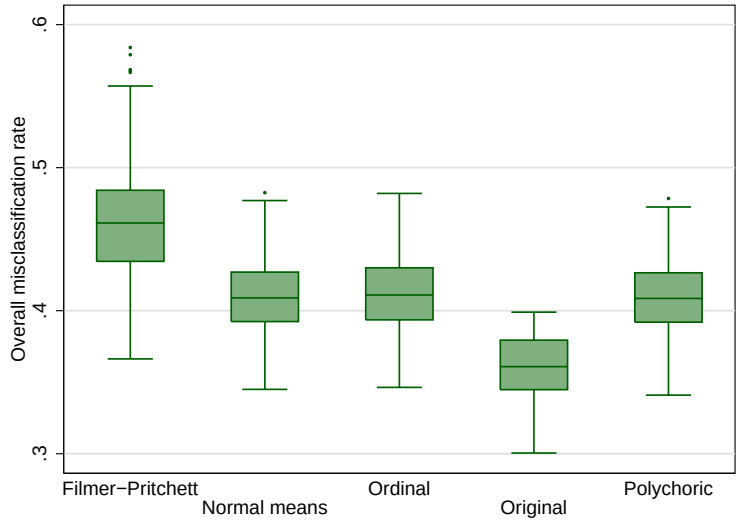
- Total number of indicators: 1–12.
- The distribution of the underlying factor: normal; uniform; lognormal; bimodal (a mixture of two normals).
- The sample sizes: 100, 500, 2000, 10000.
- The number of categories of the discrete variables: from 2 to 12.
- The proportion of the variance explained: 80%, 60%; 50%, 40%, 30%
- Various threshold α settings
- The fraction of discrete variables: from 50% (1 discrete, 1 continuous) to 100%.
- Factor loadings: all ones; some have $\lambda_k = 3$ (discrete and/or continuous)
- The analyses performed: ordinal, Filmer-Pritchett, polychoric PCA; PCA on the ordinal variables with j-th category weight set to $\mathbb{E}[x^* | x = j]$; PCA on the original continuous variables $x_1^*, \ldots, x_p^*$ (benchmark)
- Approx. 1% combinations sampled

Simulation results - 1

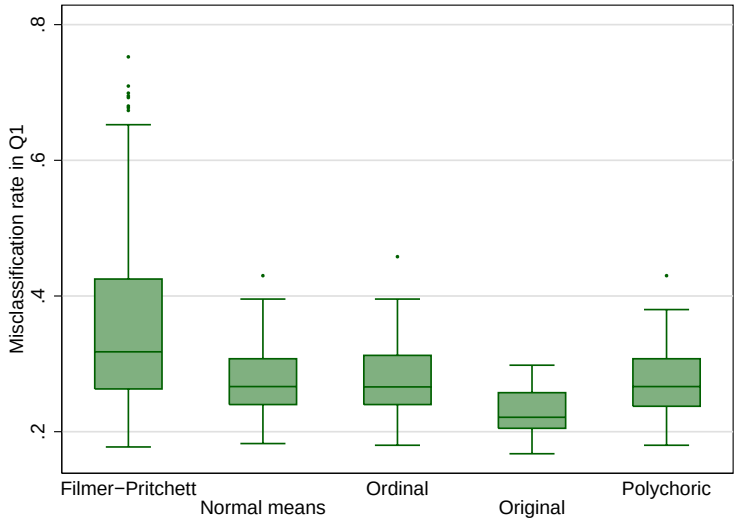
- Outcomes: (inverse probit transformation of) Spearman correlation of the empirical score with the true score; misclassification rate overall and in the 1st quintile; the reported proportion of explained variance
- Even in the most favorable situations (12 indicators, 80% of variance explained), the misclassification rates are around 24% overall (other than F-P) and 17% in the first quintile. For Filmer-Pritchett procedure, the numbers are around 29% and 27%.
- Regressions on the simulation settings give high R^2
- Filmer-Pritchett analysis is uniformly dominated by other methods
- Most important explanatory variables: S/N ratio (theoretical proportion of explained variance); heavy tails of the distribution of the underlying score; analysis type; number of variables, and their discrete/continuous character

See also Table 2 (p. 13) of Kolenikov & Angeles (2004) for regression results.

Simulation results - 2



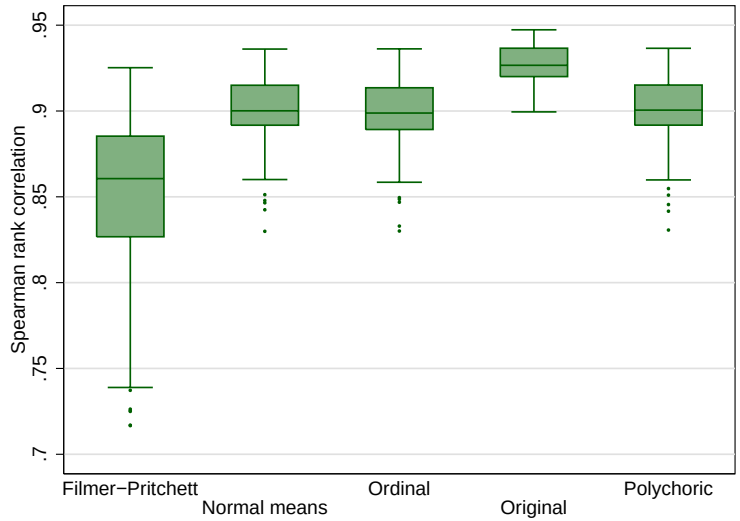
(a) Overall misclassification rate



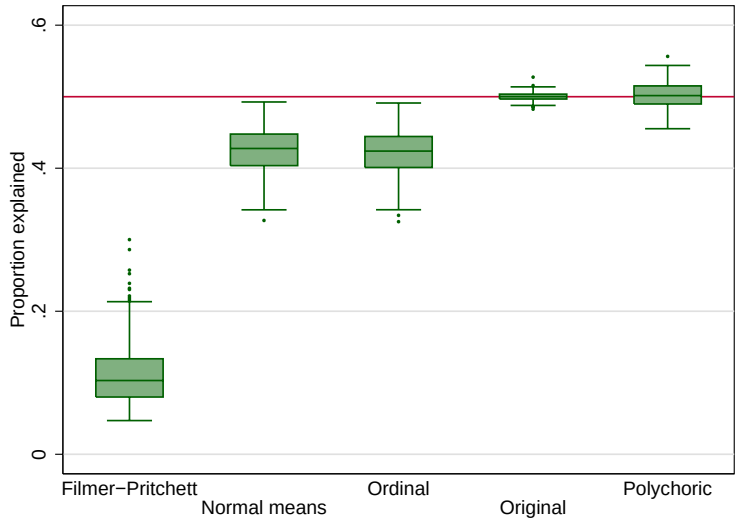
(b) Misclassification rate in Q1

Box plots for different PCA methods. Both panels (a) and (b): the lower, the better. Restrictions: 8 discrete variables, no continuous variables, sample sizes 2000 or 10000, lognormal distribution excluded, theoretical share of explained variance is 0.5 (2596 obs).

Simulation results - 3



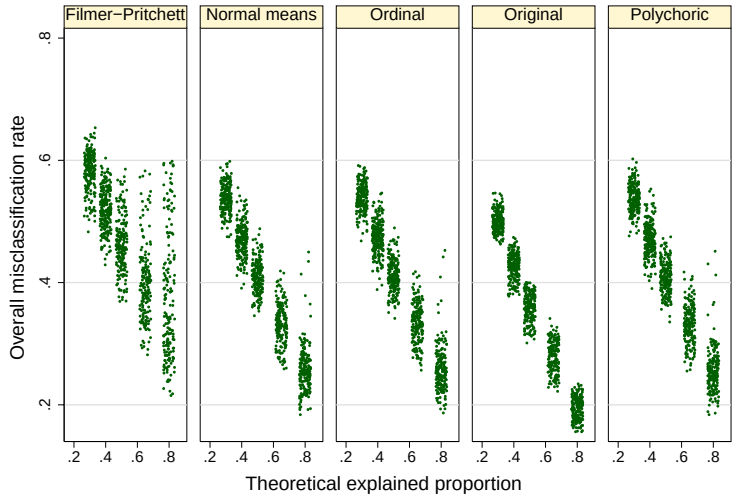
(c) Spearman's ρ between theoretical and empirical scores



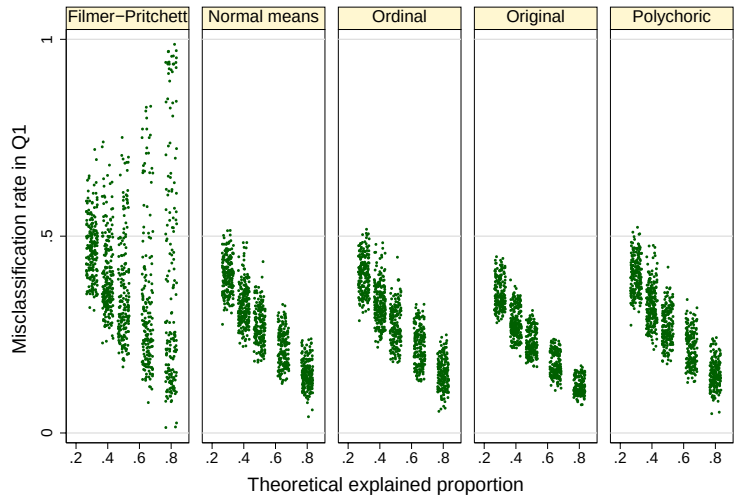
(d) Reported share of explained variance

Box plots for different PCA methods. Panel (c): the higher, the better. Panel (d): the closer to the line at 0.5, the better. Restrictions: 8 discrete variables, no continuous variables, sample sizes 2000 or 10000, lognormal distribution excluded, theoretical share of explained variance is 0.5 (2596 obs).

Simulation results - 4



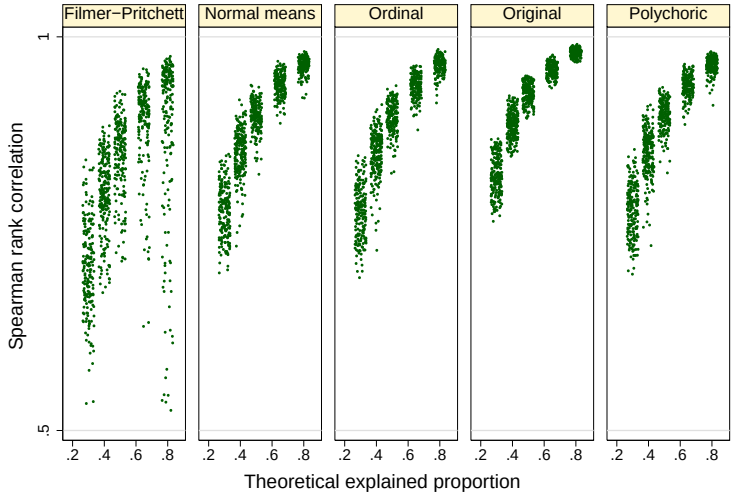
(a) Overall misclassification rate



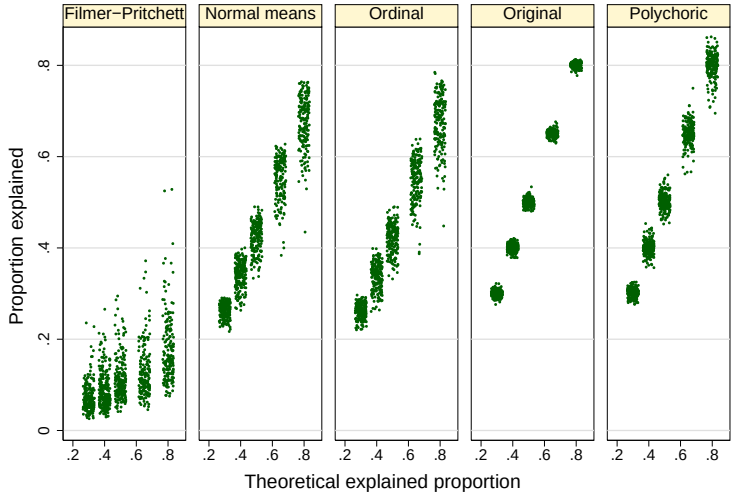
(b) Misclassification rate in Q1

Scatterplots with the underlying proportion of explained variance. Both panels (a) and (b): the lower, the better. Jitter added to show structure. Restrictions: 8 discrete variables, no continuous variables, sample sizes 2000 or 10000, lognormal distribution excluded (12880 obs).

Simulation results - 5



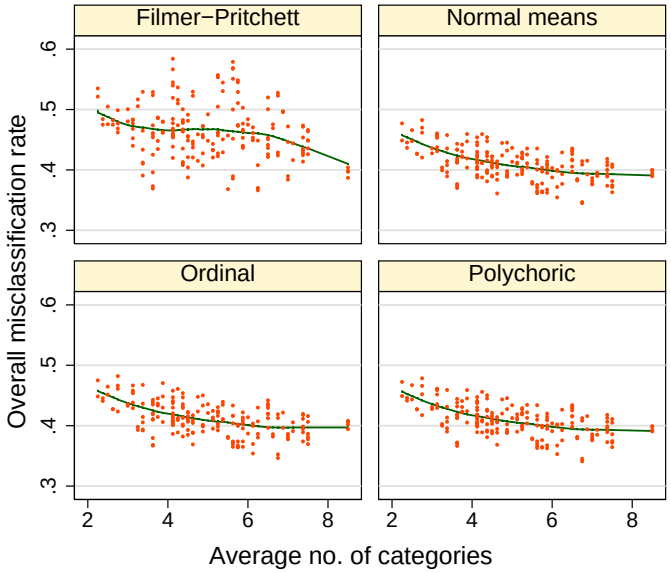
(c) Spearman's ρ between theoretical and empirical scores



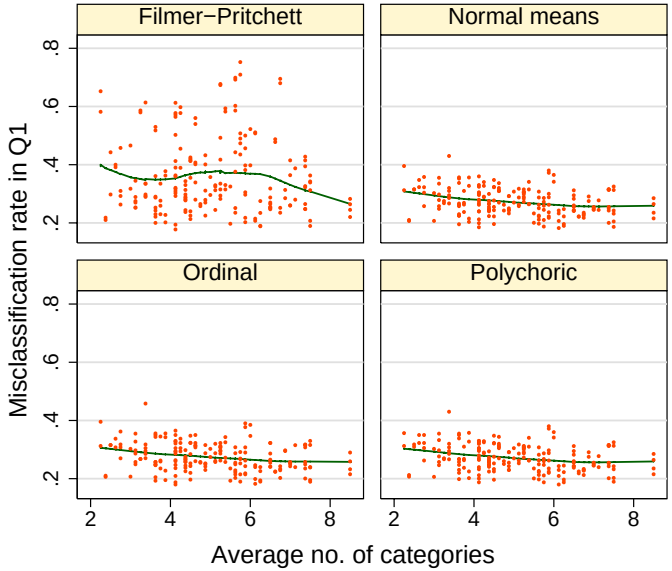
(d) Reported share of explained variance

Scatterplots with the underlying proportion of explained variance. Panel (c): the higher, the better. Panel (d): the closer to the diagonal, the better. Jitter added to show structure. Restrictions: 8 discrete variables, no continuous variables, sample sizes 2000 or 10000, lognormal distribution excluded (12880 obs).

Simulation results - 6



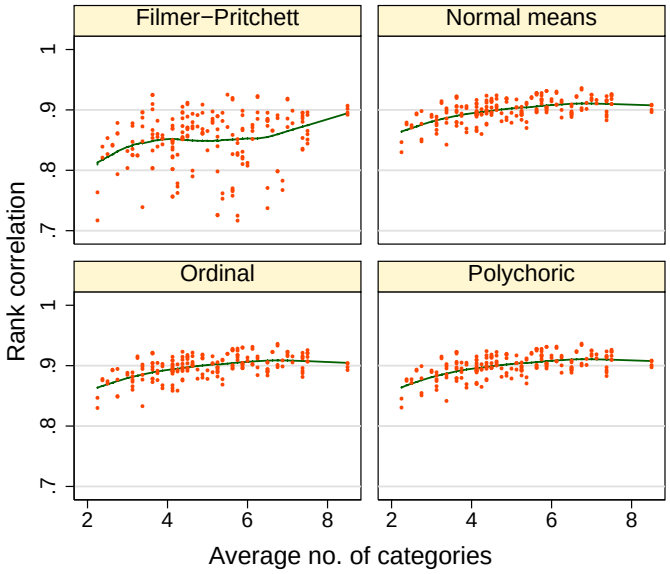
(a) Overall misclassification rate



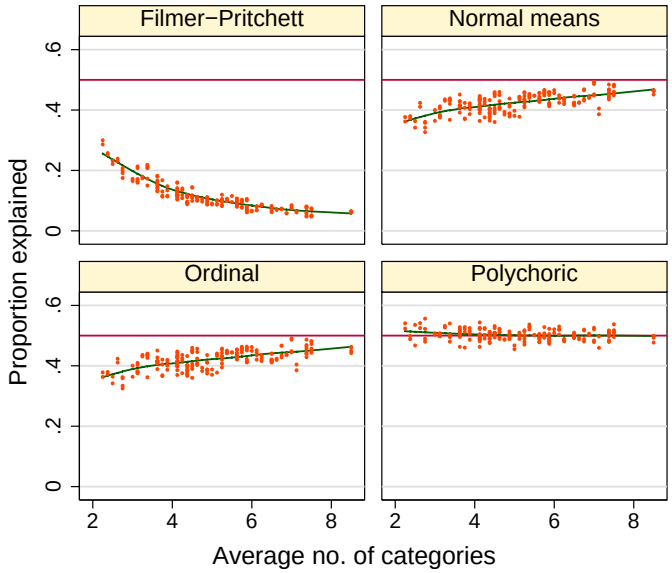
(b) Misclassification rate in Q1

Scatterplots with the average number of categories of discrete variables. Both panels (a) and (b): the lower, the better. Restrictions: 8 discrete variables, no continuous variables, sample sizes 2000 or 10000, lognormal distribution excluded, theoretical share of explained variance is 0.5 (2596 obs).

Simulation results - 7



(c) Spearman's ρ between theoretical and empirical scores



(d) Reported share of explained variance

Scatterplots with the average number of categories of discrete variables. Panel (c): the higher, the better. Panel (d): the closer to the line at 0.5, the better. Restrictions: 8 discrete variables, no continuous variables, sample sizes 2000 or 10000, lognormal distribution excluded, theoretical share of explained variance is 0.5 (2596 obs).

Conclusions

- PCA is a useful procedure in SES estimation
- Discrete data pose certain, but not major, problems
- Polychoric and ordinal scores are very similar to one another
- Only the polychoric procedure estimates the proportion of explained variance consistently
- Filmer-Pritchett procedure does not perform well when the data are ordinal
- It can be attributed to an arbitrary zero weight of the omitted category — the latter should be in the middle of the SES distribution

Further work

- More empirical examples?
- Development of appropriate factor models
- External validation, either of the single score (Bollen et al. 2002*a*), or as a part of a latent variable model (Bollen et al. 2002*b*)
- Sensitivity to misspecification in ordering of the categories for the ordinal, polychoric and group means methods, and to the choices of omitted categories, for Filmer-Pritchett procedure

References

Bollen, K. A., Glanville, J. L. & Stecklov, G. (2001), 'Socioeconomic status and class in studies of fertility and health in developing countries', *Annual Review of Sociology* **27**, 153–185.

Bollen, K. A., Glanville, J. L. & Stecklov, G. (2002*a*), 'Economic status proxies in studies of fertility in developing countries: Does the measure matter?', *Population Studies* **56**, 81–96. DOI: 10.1080/00324720213796.

Bollen, K. A., Glanville, J. L. & Stecklov, G. (2002*b*), Socioeconomic status, permanent income, and fertility: A latent variable approach, Working Paper WP-02-62, MEASURE Evaluation Project at Carolina Population Center, Chapel Hill.

Filmer, D. & Pritchett, L. (2001), 'Estimating wealth effect without expenditure data — or tears: An application to educational enrollments in states of India', *Demography* **38**, 115–132.

Hotelling, H. (1933), 'Analysis of a complex of statistical variables into principal components', *Journal of Educational Psychology* **24**, 417–441, 498–520.

Jöreskog, K. (2004), *Structural Equation Modeling With Ordinal Variables using LISREL*. Notes on LISREL 8.52. http://www.ssicentral.com/lisrel/ordinal.pdf.

Kolenikov, S. & Angeles, G. (2004), The use of discrete data in PCA: Theory, simulations, and applications to socioeconomic indices, Working paper WP-04-85, MEASURE/Evaluation project, Carolina Population Center, University of North Carolina, Chapel Hill.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (1980), *Multivariate Analysis*, Academic Press, London.

Olsson, U. (1979), 'Maximum likelihood estimation of the polychoric correlation', *Psychometrika* **44**, 443–460.

Pearson, K. (1901*a*), '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.

Pearson, K. (1901*b*), 'On lines and planes of closest fit to systems of points in space', *Philosophical Magazine* **2**, 559–572.

Rencher, A. C. (2002), *Methods of Multivariate Analysis*, John Wiley and Sons, New York.

Thomas, D. & Strauss, J. (1995), Human resources: Empirical modeling of household and family decisions, *in* 'Handbook of Development Economics', Vol. 3A, Elsevier, chapter 34.