Review of Income and Wealth
Series 55, Number 1, March 2009

SOCIOECONOMIC STATUS MEASUREMENT WITH DISCRETE PROXY VARIABLES: IS PRINCIPAL COMPONENT ANALYSIS A RELIABLE ANSWER?

by Stanislav Kolenikov*

University of Missouri

and

Gustavo Angeles

University of North Carolina

The last several years have seen a growth in the number of publications in economics that use principal component analysis (PCA) in the area of welfare studies. This paper explores the ways discrete data can be incorporated into PCA. The effects of discreteness of the observed variables on the PCA are reviewed. The statistical properties of the popular Filmer and Pritchett (2001) procedure are analyzed. The concepts of polychoric and polyserial correlations are introduced with appropriate references to the existing literature demonstrating their statistical properties. A large simulation study is carried out to compare various implementations of discrete data PCA. The simulation results show that the currently used method of running PCA on a set of dummy variables as proposed by Filmer and Pritchett (2001) can be improved upon by using procedures appropriate for discrete data, such as retaining the ordinal variables without breaking them into a set of dummy variables or using polychoric correlations. An empirical example using Bangladesh 2000 Demographic and Health Survey data helps in explaining the differences between procedures.

1. Introduction

One of the recurrent ideas and needs of development and health economics studies that use micro level data is to assess the socioeconomic status (SES) of a household or an individual. Measures of SES usually serve as inputs to another analysis such as inequality or poverty analysis, tabulation of population characteristics by quintiles or deciles, or regressions that involve SES as an explanatory or dependent variable and aim at explaining the household health status, health or economic behavior. In policy oriented applications, these measures are also utilized to make decisions regarding the allocation of projects and resources that are to benefit the poor.

Note: Ken Bollen made valuable suggestions, and Nash Herndon provided useful editorial comments. Suggestions of the two referees and the editor were crucial in putting together a concise and informative paper. Partial financial support was provided by the U.S. Agency for International Development through the MEASURE Evaluation project, Carolina Population Center, University of North Carolina at Chapel Hill, under the terms of Cooperative Agreement #GPO-A-00-03-00003-00. The views represented in the paper are those of the authors, and the remaining errors are their responsibility.

*Correspondence to: Stanislav Kolenikov, Department of Statistics, University of Missouri, Columbia, MO 65211-6100, USA (kolenikovs@missouri.edu).

Broadly speaking, socioeconomic status involves many dimensions: education and occupation of family members, their access to goods and services, and the welfare of the household as a measure of the goods and services accessibility. We shall concentrate on the economic components of socioeconomic status in this paper.

Often, straightforward numeric measures of welfare such as household income or consumption are not available or not reliable, especially in developing economies where a large fraction of economic activities may be carried out outside of the market. In such situations, the researcher has to deal with other proxies for the household wealth and/or consumption and use those in deriving an index of the household welfare. Such proxies must be easier to observe than income. Possession of durable goods and living conditions are used more and more often as those proxies. The interviewer can simply observe and record the household status, or ask sufficiently simple questions such as "Do you own a TV set?" or "What is the source of the drinking water in your house?" On one hand, those variables with a small number of clear response categories suffer fewer *reporting* errors than do income or expenditure. On the other hand, as measures of socioeconomic status, they still are subject to *measurement* error, since they are imperfect measures of SES.

The use of a single proxy is likely to lead to unreliable and/or unstable results, so a natural idea would be to incorporate a number of such proxies to compensate for various measurement errors. Section 2 reviews the variables and the methods typically used. Usually between 10 and 20 characteristics can be observed, and then the analyst must have a method for aggregating such proxies. By far the most popular method is to assign coefficients, or weights, to those observed variables, and sum them up. The weights may come from some economic considerations, such as assigning a monetary value for durable goods; from statistical considerations, such as principal component analysis (PCA); or from other considerations, such as putting all coefficients to one.

Principal component analysis is a standard multivariate technique developed in the early 20th century (Pearson, 1901b; Hotelling, 1933) in psychometrics and multivariate statistical analysis for similar purposes of aggregating information scattered in many numeric measures, such as student scores on several tests. It is described in many multivariate and dedicated textbooks such as Anderson (2003), Mardia *et al.* (1980), Flury (1988), Jolliffe (2002) and Rencher (2002). In economics, the method has been applied to the studies of cointegration and spatial convergence (Harris, 1997; Drakos, 2002), development (Caudill *et al.*, 2000), panel data (Bai, 1993; Reichlin, 2002), forecasting (Stock and Watson, 2002), simultaneous equations (Choi, 2002) and economics of education (Webster, 2001). Krelle (1997) gives a review of a number of methods aimed at estimation of unobservable variables, including PCA.

In their recent work, Filmer and Pritchett (1998, 2001) used PCA to construct socioeconomic indices. They used the data on household assets (the most important durable goods such as clock, bicycle, radio, television, sewing machine, motorcycle, refrigerator, car), type of access to hygienic facilities (sources of drinking water, types of toilet), number of rooms in dwelling, and construction materials used in the dwelling. The methodology was quickly picked up by the World Bank (Gwatkin *et al.*, 2003a, 2003b, 2007) and Demographic and Health

Surveys (DHS)[1] as the way to assess socioeconomic status of a household based on the household assets and facilities. The DHS are among the richest, most reliable and representative sources of data for health and demographic analysis in developing countries. At the time of publication, there were 76 countries where the standard DHS surveys were implemented, and data collected at least once. The DHS, however, have a significant limitation, which is that they do not collect data on income or consumption. The PCA methodology as proposed by Filmer and Pritchett has indeed become very popular because it tackles that data constrain and therefore, it allows the researchers to expand the use of DHS data to analysis of inequities in health and to include SES proxies in analysis of determinants of health.

PCA, however, was originally developed for multivariate normal data, and is best used with continuous data. Its application by Filmer and Pritchett was an ad hoc solution to the problem of measuring wealth effects in the absence of consumption data. They acknowledged that justification of PCA or study of its properties was limited, and that is the concern this paper seeks to address.

The central section addressing those issues is Section 3, where the main problems arising from the discrete nature of the data are reviewed, and statistical properties of the Filmer–Pritchett procedure are derived. Unlike with continuous data where PCA is straightforward, there are multiple possible implementations of PCA in the case of discrete data. We propose several procedures and examine their performance in settings typically available in household health surveys for measuring SES. We designed and carried out a large simulation project which also confirmed that the Filmer–Pritchett procedure can be improved upon, particularly when dealing with ordinal variables. The simulation results are reported in Section 4, and a smaller study of sensitivity for the main assumptions of the new methods is given in Section 5. Finally, an empirical illustration with Bangladesh Demographic and Health Survey data is given in Section 6; Section 7 concludes.

2. Proxy Measurement of Socioeconomic Status

In this section, we review the main approaches to measurement of socioeconomic status with concurrent proxy variables in applications typical for health economics. We give both theoretical arguments about what SES is, and practical considerations based on the variables typically available in demographic and health surveys.

There are many concepts of socioeconomic status. One such view is that SES is essentially a univariate concept. From this perspective, there is a single fundamental dimension that underlies SES. An example of such an approach focusing on economic status is Friedman's permanent income hypothesis (Friedman, 1957). Friedman argued that income is composed of two components: permanent and transitory. The former is largely determined by the human and production capital of the household or individual, as well as its wealth, while the latter is affected by the market fluctuations and other random occurrences. In any given period, the observed income is the sum of the two. Also, Friedman argued that consumption

[1]See http://www.measuredhs.com.

expenditure is largely driven by the permanent income, at least when the economic agents can borrow against future incomes.

This theory directly provides for measurement of economic status by income or consumption expenditure over a specific time interval, such as a month or a year. The longer the observation period, the more accurate the measure of the (permanent) income. Also, collecting expenditure data is more straightforward and reliable, while income data are often unreliable and difficult to collect in developing countries (Hentschel and Lanjouw, 1996). Therefore, annual household expenditures may provide better permanent income proxies of longer-term economic status (Deaton, 1992).

There are some disadvantages to this class of measures. First, many surveys do not collect information on expenditures because of the time and cost involved. Second, in developing nations, households may not be able to smooth consumption behavior over time by borrowing and saving. Third, studies have shown that measures of consumption can also be error prone (Scott and Amenuvegbe, 1990; Bouis, 1994). This is particularly true for analyses of developing countries where income and expenditure data are often of poor quality.

Other approaches to SES measurement highlight separate dimensions of social stratification and predict that different dimensions can have different consequences. Bollen *et al.* (2001) reviewed the use of SES and class in child-health and fertility studies. They found that a more common assumption in the literature was that SES is composed of distinct components, each capable of exerting separate effects.

Since economic status and permanent income are theoretical concepts that are not directly measurable, a wide range of proxy measures have been proposed. They differ according to whether they focus on measurement of observed stocks or assets, or on the resource flows over of a chosen time period, or whether they attempt to capture economic status indirectly through occupation, education, or other related measures (Bollen *et al.*, 2001).

When income and expenditure data are not used, measures of households' ownership of consumer durable goods and housing quality are frequently employed to capture household economic status, as they are easier to collect than either income or expenditure data. Using these data, an applied researcher can select a single variable to proxy economic status, or construct one or more indices based on a composite of different factors with potentially equal or variable weights.

A review conducted by Angeles and You (2007) summarized information on the number and type of variables in DHS that could be used to calculate SES indexes. The study included 76 country/year household surveys during the past 13 years (1994–2007). Results show that two major categories of SES variables were included in DHS: housing characteristics and possession of durable goods. Housing characteristics variables that are available in most surveys include type of drinking water, type of toilet facility, main floor material, sources of cooking fuel, whether the household has electricity, and the number of bedrooms. Durable goods variables that are available in most surveys include whether the household has radio, TV, refrigerator, bicycle, motorcycle, car, and telephone. Some surveys also include countryspecific characteristics, such as whether the household has livestock, a dacha, a boat, a bednet, etc.

The average number of variables that could be used to calculate SES index is 20, with a range from 11 to 42. The average number of binary variables that could be used is 12, ranging from 5 to 32. The average number of categorical variables is 6, ranging from 3 to 17; the average number of categories is 7, with a range of categories from 3 to 16. There were on average 2 variables that were truly quantitative, either continuous (time to get to a source of drinking water for instance) or count (number of rooms typically.) In most household surveys, the majority of variables used to calculate PCA are binary variables; on average about 60 percent of variables are binary, the largest percentage is 75 percent (Mali DHS conducted in 2001). There is only one survey (Bolivia DHS conducted in 1998) that has fewer binary variables than categorical variables (out of 21 variables that could be used to calculate SES index, 8 (38 percent) are binary variables, 10 (48 percent) are categorical variables).

An extensive review of the existing methods of SES assessment in application to the health and fertility studies is given in Bollen *et al.* (2001, 2002). They note that "measures of SES . . . vary widely within and between disciplines regardless of the outcome," and "empirical implementations of SES . . . are often driven by data availability and the empirical performance of indicators as much as they are by theoretical groundwork." They compare the performance in terms of external validity, defined there as the explanatory power in a regression set up as follows. The dependent variable is fertility (whether the woman had any births in the last three years), and explanatory variables are demographic controls and one of the SES measures: (i) a simple sum of the assets, i.e. total number of durable goods possessed by the household; (ii) sum of current values of those durable goods, as assessed by the household itself; (iii) sum of median values of the durable goods possessed by the household, where the median value of the asset across all households is taken as the market price of an item; (iv) principal component scores; and also measures based on single variables such as occupational prestige, or expenditure per adult. They found that the best fitting measures were the principal component measure and simple sum of asset indicators.

Filmer and Pritchett (2001) adopted a version of principal component analysis procedure where the variables with multiple categories, such as the source of water or materials used in dwelling construction, are broken down into a set of dummy variables, as is typical in regression analysis. They performed PCA on the resulting set of binary indicators, and used the first principal component as a measure of SES.

Vyas and Kumaranayake (2006) reviewed issues related to choice of variables and data preparation, and problems such as data clustering were addressed. Their study used data from Brazil and Ethiopia DHS, including whether a household has electricity, radio, television, refrigerator, car, bicycle or telephone; and the number of rooms for sleeping, source of water supply, type of sanitation facility, and type of floor material. Of 11 variables, 3 were categorical, with number of categories ranging from 7 to 9.

A series of World Bank reports (Gwatkin *et al.*, 2007) provides basic information about health, nutrition, and population (HNP) inequalities in 56 developing countries. Using DHS, the reports present data about HNP status, service use, and related matters among individuals belonging to different socioeconomic

classes. The principal focus is on differences among groups of individuals defined in terms of the wealth or assets of the households. Wealth index scores were constructed using the Filmer–Pritchett procedure. The following are some examples of asset variables used in constructing the wealth index. In the 2004 Bangladesh report, 20 asset variables were included: availability of electricity, radio, television, bicycle, motorcycle, telephone, almirah, table, chair, clock, bed, sewing machine, land, a domestic worker, water source, type of toilet, type of floor, type of wall, type of roof, and type of cooking fuel. Among them, 6 are categorical and the average number of categories is 5, ranging from 3 to 7. In the 2004 Colombia report, 30 asset variables were included, among which 7 are categorical and the average number of categories is 8, ranging from 5 to 12. In the 2003 Ghana report, 17 asset variables were included, among which 5 are categorical and the average number of categories is 7, ranging from 5 to 11.

Thomas (2007) examined child mortality and socioeconomic status among migrants and non-migrants based on micro data from the 3 percent sample of the 1996 population census of South Africa, with 5 ordinal and 2 binary variables. Hong and Hong (2007) used the 2000 Cambodia DHS to examine how household and community economic inequalities affect nutritional status in women, with 6 categorical variables and 14 binary variables. Sumarto *et al.* (2007) used wealth index and principal components analysis to predict consumption expenditure and poverty using non-consumption indicators. Fernald (2007) explored the associations of body mass index (BMI), SES and beverage consumption in very low-income Mexican adults. This study used education, occupation, household income, and housing/assets characteristics to construct SES. Proxy measures of objective SES were generated: one summary measure of household assets (12 variables) and the other of housing quality (6 variables). Deressa *et al.* (2007) assessed household and socioeconomic factors associated with childhood febrile illnesses and treatment seeking behavior using a study that was conducted in Adami Tutu district in Ethiopia during the peak malaria transmission season in 2003.

3. PCA and Related Procedures

Given a set of variables $y_1, \ldots, y_p$, the principal component analysis seeks to find the linear combinations of those variables with maximum variance:

$$\frac{\mathbb{V}[\mathbf{a}'\mathbf{y}]}{\|\mathbf{a}\|} \to \max_{\mathbf{a}}.$$

The solutions are given by the eigenvalues and eigenvectors of the covariance matrix of y. The main principles of principal component analysis are covered in Appendix A, alongside all other appendices, available externally on the Internet.[2] For a more extensive introduction to the topic, readers not familiar with geometric and statistical properties of PCA might wish to consult Mardia *et al.* (1980), Flury (1988), Jolliffe (2002), or Rencher (2002).

[2]The appendices to this paper are available at http://web.missouri.edu/~kolenikovs/papers/roiw-309-appendices.pdf. The appendices are also available at the following Wiley-Blackwell website: http://dx.doi.org/10.1111/j.1475-4991.2008.00309.x.

An intuition underlying PCA is that there is one or few variables that underlie all of the structure in (covariance of) the data. A related, although distinct, multivariate model that makes explicit use of the underlying latent variables (such as welfare) is the *confirmatory factor analysis* (CFA) model (Bartholomew and Knott, 1999):

$$(3.1) \qquad \mathbf{y} = \Lambda\xi + \delta, \quad \mathbb{E}[\delta] = 0, \quad \mathbb{V}[\delta] = \Theta_\delta = \mathrm{diag}[\delta_1, \ldots, \delta_K], \quad \mathbb{V}[\xi] = \phi,$$
$$\delta, \mathbf{y} \text{ and } \Lambda \text{ are } K \times 1 \text{ vectors}$$

where ξ is the latent factor, and $\mathrm{y} = (y_1, \ldots, y_K)$ are observed variables. In a number of circumstances, such as equal values of the coefficients λ_k and variances 0_{kk}, the two methods give identical answers. The model (3.1) will be used for simulations in Section 4.

In Section 3.1, we discuss the problems with discrete data, and in Section 3.2, introduce polychoric and polyserial correlations as alternative estimators of the correlation between discrete variables (or rather between their latent continuous counterparts). The Filmer–Pritchett procedure is discussed in Section 3.3, and nominal categorical variables are treated in Section 3.4.

3.1. *Discrete Data*

One of the assumptions underlying the "classic" formulation of the principal components is that the input variables are multivariate normal, or at least that normality is a reasonable distributional approximation. When the data are discrete, as happens in practice with data such as DHS, this assumption is clearly violated. There are several kinds of discrete data one can encounter in empirical analysis (Hand, 2004). Most often the discrete data are *binary*, i.e. a variable that can only take one of two values, such as gender (male/female), ownership of a car, or a decision to participate in a program. If there are several categories of a discrete variable, they may or may not have some natural ordering. If they do, the discrete data are referred to as *ordinal*: there are several categories with a monotone relation among them. The examples might be: different levels of education (no education, primary, secondary, higher, professional or advanced degree), subjective well-being on a ladder/Likert scale (from 1 to 9 where 1 is the most miserable person, and 9 is the happiest one), or different construction materials used in the building (no roof, a straw roof, a wooden roof, a tile roof). Often, binary data can be viewed as a special case of ordinal data with only two categories (having a car is better than not having one). There may be no particular order of the categories for other types of *nominal* categorical variables, such as race and gender of a person, industry of a firm, or a geographical region. Yet another type of discrete data is *count* data, such as the number of crimes in a given area in a year, or a number of children born to a woman.

There are numerous implications of the discrete character of the data if the observed discrete y_k's are used directly in the standard principal component analysis. The problems related to the discrete data have received a considerable attention in social measurement literature (Olsson, 1979; Bollen and Barb, 1981; Johnson and Creech, 1983; Babakus *et al.*, 1987; Dolan, 1994; DiStefano, 2002).

Obviously, the discrete data violate distributional assumptions in methods where continuous variables are assumed or expected. Also, even despite the finite range, the discrete data tend to have high skewness and kurtosis, especially if the majority of the data points are concentrated in a single category. Example 1 of Appendix B considers two variables, with four and three categories respectively, and finds that they have skewness of –0.40 and 0.39, and kurtosis of 2.52 and 2.04.

3.2. *Polychoric and Polyserial Correlations*

There is a substantial body of literature on the use of discrete data in multivariate methods. Interestingly, the history of both discrete data methods and multivariate methods goes back to the same person. The early versions of the principal component analysis were introduced in Pearson (1901b), while Pearson (1901a) introduced *tetrachoric* correlation for a two-by-two contingency table as an improved measure of correlation between two binary variables. Further work with major contributions of Pearson and Pearson (1922) and Olsson (1979) introduced concepts of *polychoric* and *polyserial* correlations as the maximum likelihood estimates of the correlation between the unobserved normally distributed continuous index variables underlying their discretized versions. Bollen and Barb (1981), Babakus *et al.* (1987), Dolan (1994), and DiStefano (2002), among others, have looked at the effects of categorization in a closely related area of structural equation modeling with latent variables, also known as linear structural relations. A general treatment of problems with multivariate ordinal data is given in Jöreskog (2004). So far, applications of the polychoric correlations in economics publications are extremely sparse (Di Bartolo, 2000), and the method is beginning to gain recognition in public health studies (Medina-Solís *et al.*, 2006).

A typical framework of analyzing ordinal data is a multivariate extension of the threshold structure of the standard ordinal probit model (Maddala, 1983; Wooldridge, 2002). If the observed y_k's are ordinal with the categories 1, . . . , d_k, then it is assumed that they are obtained by discretizing the underlying y_k^* according to the set of thresholds $\alpha_{k1}, \ldots, \alpha_{k,d_k-1}$:

$$y_k = r \quad \text{if } \alpha_{k,r-1} < y_k^* < \alpha_{k,r} \tag{3.2}$$

where $\alpha_{k,0} = -\infty$, $\alpha_{k,d_k} = +\infty$. Example 1 in Appendix B demonstrates the concept. It is possible to recover the correlation between the underlying continuous starred variables using their discrete manifestations. Suppose two ordinal variables y_1, y_2 are obtained by categorizing two variables y_1^*, y_2^* with distribution

$$\begin{pmatrix} y_1^* \\ y_2^* \end{pmatrix} \sim N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \quad -1 \le \rho \le 1. \tag{3.3}$$

The categorizing thresholds for the two variables are given by $\alpha_{1,0} = -\infty < \alpha_{1,1} < \ldots < \alpha_{1,d_1-1} < \alpha_{1,d_1} = \infty$, $\alpha_{2,0} = -\infty < \alpha_{2,1} < \ldots < \alpha_{2,d_2-1} < \alpha_{2,d_2} = \infty$, so that $y_i = k$ when $\alpha_{i,k-1} < y_i^* \le \alpha_{i,k}$, $i = 1, 2$. Then the theoretical proportions $\pi(m, l;$

ρ, α) of the data in the cell (m, l) of the contingency table are given by (B.2) in Appendix B. Assuming that observations are i.i.d., the likelihood can be written down as

$$L(\rho\alpha) = \prod_{i=1}^{N}\prod_{m=1}^{d_1}\prod_{l=1}^{d_2}\pi(m, l; \rho, \alpha)^{I(x_{i,1}=m, x_{2,i}=l)} = \prod_{i=1}^{N}\left[\pi\left(y_{i,1}, y_{i,2}, \rho, \alpha\right)\right] \tag{3.4}$$

$$\ln L = \sum_{i=1}^{N}\ln\pi\left(y_{i,1}, y_{i,2}; \rho, \alpha\right). \tag{3.5}$$

Maximizing over ρ and α, one obtains the *polychoric correlation* of y_1 and y_2. Being the maximum likelihood estimate, it is consistent, asymptotically normal, and asymptotically efficient, as the regularity conditions for those properties can be verified to hold. In moderate size samples ($n = 500$), Olsson (1979) found the polychoric estimates to have slight upward bias.

In practice, the estimation is performed in three stages. First, the thresholds are estimated from the marginal distribution of x_i:

$$\hat{\alpha}_{i,j} = \Phi^{-1}\left(\frac{-1/2 + \#\{x_i \le j\}}{N}\right), \quad j = 1, \ldots, d_i, \tag{3.6}$$

Second, the correlation coefficient is estimated by maximizing (3.5) conditional on α. This procedure does not yield the maximum likelihood estimates. However, Olsson (1979) found in his simulations that the differences are below $0.5 \cdot 10^{-2}$ (cf. the standard error of 0.02–0.05 in his sample sizes of 500), and explains that the difference is due to correlation between $\hat{\rho}$ (the correlation estimate itself) and $\hat{\alpha}$ (the set of thresholds). Those correlations are zero when $\rho = 0$, and rise to about 0.2 when $\rho = 0.85$. Maydeu-Olivares (2001) derives the distribution of the estimates of the polychoric correlation from the two-stage procedure, and also finds that the discrepancies between the two methods are negligible. Third, the estimates are combined into an estimate of the correlation matrix.

Note that in the framework of the model (3.1), y's are *dependent* variables, and if ξ were observed, we would use ordered dependent variable models to analyze the relations between ξ and y.

If we are computing the correlation between a discrete and a continuous variable, then a correction that works in the same way as the polychoric correlation is the *polyserial* correlation. The likelihood for the discrete variable y_1 with underlying standard normal y_1^* discretized according to the thresholds $\alpha_{1,0} = \infty < \alpha_{1,1} < \ldots < \alpha_{d_1} < \alpha_{1,K} = \infty$, and the continuous variable y_2 (assumed WLOG to have the standard normal distribution) is:

$$\begin{aligned} L(\rho, \alpha; y_1 = k, y_2) = f(y_1 = k, y_2; \rho, \alpha) &= \mathrm{Prob}\left[\alpha_{1,k-1} < y_1^* \le \alpha_{1,k} \,|\, y_2\right]\phi(y_2) \\ &= \left(\Phi\left(\alpha_{1,k} - \rho y_2\right) - \Phi\left(\alpha_{1,k-1} - \rho y_2\right)\right)\phi(y_2) \end{aligned} \tag{3.7}$$

since $\mathbb{E}[y_1^*|y_2] = \rho y_2$. Assuming independence of observations to sum up the log-likelihood, the resulting expression can be maximized (jointly or in two stages) with respect to α and ρ to give the polyserial correlation between the two variables.

In the multivariate setting with more than two variables, the estimate of the overall correlation matrix is obtained by combining the pairwise estimates of the polychoric, polyserial, or moment correlations. Then one proceeds to the PCA in the standard manner, i.e. by solving the eigenproblem for the estimated correlation matrix.

It is not guaranteed that the correlation matrix obtained in this manner will be non-negative definite. However, this is not an obstacle for PCA. As all individual correlation estimates are consistent for their population counterparts, the matrix as a whole is consistent for the population correlation matrix (of underlying continuous starred versions). If the resulting correlation matrix is not positive definite, then non-positive entries of it are solely due to sampling fluctuations, and have to be small in magnitude. As our primary interest is in the largest eigenvalue(s), the fluctuations in the smaller ones are not worrisome.

A consequence of the discretization (3.2) is that the covariances or correlations between the observed variables are not equal to the "true" covariances or correlations of the (unobserved) underlying index variables. Example 2 of Appendix B shows that in the extreme case of dichotomizing the continuous distribution, the correlations are always underestimated. In a more general case of more than two categories, categorization can be viewed as a measurement error with non-linear properties, and the authors are not aware of research showing that correlations go down because of discretization. It is, however, very well established empirically in sociological literature cited above. As long as the covariance matrix is estimated with bias, the principal components may not be estimating the underlying welfare score accurately. Since the correlations of the discretized variables are smaller than those of the underlying scores, the proportion of explained variance is also going to be lower for the PCA based on discrete variables as compared to the PCA based on the (unobserved) underlying continuous variables.

An option that seems to lie between the polychoric correlation analysis and the analysis based on the ordinal indicators is to estimate the mean of the underlying normal variable y^* conditional on a particular category of the observed ordinal indicator $y = j$:

$$\mathbb{E}[y^*|y = j] = \int_{\alpha_{j-1}}^{\alpha_j} u\phi(u)\,du = \phi(\alpha_{j-1}) - \phi(\alpha_j), \quad \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}. \tag{3.8}$$

This value can be used instead of $y = j$ to make the variable less skewed and/or kurtotic, as well as to make the distance between the categories more informative, rather than assuming the distance between categories 1 and 2 to be the same as the distance between categories 2 and 3, or 3 and 4.

If analysis of polychoric and polyserial correlations is difficult or impossible, a crude plug-in strategy would be to use the discrete x's as if they were continuous in the PCA. Continuing our analogy with the econometric models, this is what

would happen if one runs OLS instead of ordered logit/probit on the ordinal data.[3] If the ordinal data are used as if they were continuous, problems may arise. The violations of the distributional assumptions in PCA incurred by ordinal data are the same sort of violations that econometricians are concerned with in the discrete dependent variable models, such as the logit/probit models and their ordered versions. Indeed, within the framework of the model (A.5), the observed indicators are actually the dependent variables.

3.3. *The Filmer–Pritchett Procedure*

A modification of PCA popular in development economics is due to Filmer and Pritchett (2001). They proposed to create a dummy variable for each category of the discrete variable, so the variable "Source of drinking water" with categories 1 for lake or stream, 2 for tube well, 3 for pipe outside the dwelling, and 4 for the pipe inside the dwelling will be represented by four dummies (or three if a perfect collinearity is to be avoided). The reasoning behind this proposal may have been the common recommendation to use individual binary indicators whenever the categorical variable is to be used in regression analysis. The recommendation is certainly warranted when the variable is an explanatory one. For the purposes of PCA, however, we want to stress that the input variables should be treated as *dependent* ones. The variability in assets is caused by variability of welfare, rather than vice versa.

The use of dummy variables in PCA introduces spurious correlations. The dummy variables produced from the same factor are negatively correlated, although the strength of dependence declines with the number of categories. The PCA method then needs to take into account both the fundamental (usually positive) correlations between observed variables and the spurious (negative) correlations between the dummy variables produced from a single factor. If the measurement error contained in proxy indicators of SES is large, then the correlations implied by the common SES may be comparable to the spurious correlations due to the common categorical origin. The PCA procedure may not be able to recover the SES from the data, as the directions of greater variability may now correspond to those spurious correlations. The goodness of fit measures such as proportion of explained variance are going to deteriorate, too: even if the first component is reproduced adequately, the denominator of the explained variance contains more terms accounting for more variables created for the analysis.

Also, the Filmer–Pritchett procedure loses all of the ordinal information, if there were any. It can be argued that one of the strengths of the Filmer–Pritchett method is that it does not make any assumptions regarding the ordering of the categories. However, additional information brings higher efficiency. Here, the

[3]In the case of PCA, one can find an additional justification for this approach by noting that computing correlations between ordered categories can be viewed as computing Spearman's rank correlation ρ_S instead of Pearson's moment correlation. Then, to be consistent, one should compute Spearman's ρ_S for each pair of variables, and use the matrix of rank correlations to run PCA on (Lebart *et al.*, 1984, Section I.3.4). The rank correlations are robust to non-normality of the variables, which is important for both the discrete data, and the income data which are usually heavily skewed and/or affected by outliers.

weakest form of the model assumptions used is that the researcher can provide ordering of categories based on the substantive knowledge of the problem. Our simulation analysis in Section 5 shows that this "assumption-free" argument in favor of the Filmer–Pritchett procedure is not warranted, as performance of the Filmer–Pritchett procedure is not universally better when the ordering of the categories is misspecified.

Two analytical examples are worked out in Appendix B. Example 3 shows some of the possible consequences of discretization. Suppose a researcher has a single ordinal variable as a measure of SES. If binary indicators are used for each of the categories, then those variables have negative cross-correlations. The proportion of the explained variance does not show that all the data came from a single factor, and all of the variation could be explained with a single score. If some categories are approximately equally populated, then the principal components are not well defined. In the extreme case of all categories having the same proportions, any weights that sum up to zero may serve as valid first principal component coefficients. The empirical implication of this is that the first principal component will be highly unstable and wiggle due to sampling fluctuations that would make some categories more populated.

The algebra is further developed in Example 4 which derives a more explicit solution for the case of three categories. It also shows that the first principal component gives the largest weight to the dummy variable of the category with the largest number of observations, and the second largest weight of opposite sign, to the second largest category. Indeed, those are the variables that have the largest variability, and define the largest off-diagonal entry of the correlation matrix. This is also a general result supported by empirical evidence on data sets with dummy variables: the first principal component would tend to connect the most populated categories, and the following components would try to add the next most populated ones. Hence, in this case PCA capitalizes on the measurement error introduced by the categories of observed housing materials, toilet type, etc., but not on the substantive ordering of households from poor to rich.

As long as the natural ordering of categories is not generally reproduced by the principal component analysis, the only condition that identifies the ordering could be the use of monotone variables for which the higher values really mean higher SES. The continuous variables such as income, expenditure, value of the property, etc., will serve best. The binary ownership indicators tend to produce reasonable results in practice, too. Otherwise, unless the two largest categories are the poorest and the richest members of the population by chance or by design, the first principal component would fail to give a meaningful direction of the welfare gradient.

To summarize, in absence of anchoring ordinal information, the Filmer–Pritchett procedure only pays attention to the number of individuals in each category, rather than to their relative standing on an SES scale. Spurious negative correlations between the dummy variables produced from the same categorical variable are introduced, and hence the principal components procedure will find more "significant" (or "interesting") components than there really are in the data.

3.4. *Categorical Non-Ordinal Variables*

In some cases, a researcher may be faced with a handful of categorical variables that do not have any obvious ordering. Such variables might represent gender, ethnicity, or religion of an individual, or geographical regions where the person or the household resides. It might be noted that most of the time those variables are not the outcome variables, unlike the property ownership. Rather, they would be more likely viewed as explanatory rather than dependent variables in any reasonable analysis. Thus, to provide a framework for their inclusion, one must have a model with a multivariate explanatory variables vector, the unobserved SES, and observed proxies of SES. Such models are known as *MIMIC* (multiple indicators and multiple causes) models. The explanatory variables in those models are also called causal indicators (Bollen, 1984; Fayers and Hand, 2002) or formative indicators (Diamantopoulos and Winklhofer, 2001), as opposed to effect indicators or reflective indicators, which are thought of as functions of the underlying factors.

In terms of model set up, the basic equation (3.1) will need to be extended to allow explanatory variables x that affect SES:

$$\begin{aligned} \xi &= \beta' x + \varepsilon, \\ y_k &= \Lambda_k \xi + \delta_k, \quad k = 1, \ldots, p. \end{aligned} \tag{3.9}$$

This model is a special case of a structural equation model with latent variables (Bollen, 1989; Bollen and Long, 1993; Bartholomew and Knott, 1999; Kaplan, 2000).

A thorough introduction to the topic of latent variable modeling highlighting socioeconomic status applications is given by Bollen *et al.* (2006). They argue that a good measure of SES should possess some external validity. In other words, if there is a theory predicting certain outcomes, such as health behaviors, then the better the measure of SES, the stronger it will show in the model explaining that behavior. They compared a number of different approaches to SES measurement, and found that the latent variable modeling approach provides the most clearly seen effects of permanent income on fertility. We use a similar approach in Section 6.6 where an empirical demonstration is used to compare the performance of various measures of SES.

Treatment of structural equation models is available in SAS PROC CALIS, as well as by GLLAMM add-on to Stata (Rabe-Hesketh *et al.*, 2005). There is also a number of dedicated packages for structural equation modeling, such as M*plus*, AMOS (a part of SPSS), EQS, LISREL, and others. For instance, M*plus* software user's guide (Muthén and Muthén, 2004) provides syntax for a MIMIC model in Example 5.8. More details on estimation and prediction using MIMIC model using GLLAMM are given in Appendix C.

If such software is not available, the researcher is facing a difficult choice. One of the referees of this paper suggested several options:

(a) If truly categorical variables are to be included, i.e. there is no meaningful way to order them, then do not use PCA to measure the SES.

(b) Exclude truly categorical variables. (It could even be argued that, as proxies for SES, categorical variables only make sense if they can be ordered.)
(c) Force an ordering on the categorical variables and then treat them as ordinal.
(d) Define a separate SES for each category, if there are very few categories.
(e) Use the proposal in this paper for ordinal variables, and dummies for the categorical variables.

It is shown in Appendix C that empirical Bayes prediction of ξ is given by a certain linear combination of x and y. As the x variables are used in explanatory ones, the analogy with linear regression is now valid, and breaking them into dummy variables is more solidly justified than for the response y variables. If only PCA tools are available, we gravitate towards option (e). The significance of the explanatory categorical variables can be assessed by running PCA without categorical variables, as in option (b), and then performing ANOVA using the resulting SES measure as the dependent variable. A non-significant result would imply that the categorical variables do not contribute to SES in the current application.

Finally, an SES index might be constructed using canonical correlations. While PCA aims at finding the linear combinations with greatest variance based on a set of variables, canonical correlations start with two sets of variables and aim at finding the linear combinations such that the correlations between two resulting indices will be maximized. The canonical correlation procedure can then be used as follows: (i) all the explanatory categorical variables are put into the first group of variables in their dummy format; (ii) all other SES proxies are put into the second group of variables; (iii) an appropriate correlation matrix is obtained; (iv) canonical correlation analysis is performed on that matrix; and (v) the linear combinations of both explanatory and proxy variables corresponding to the greatest canonical correlation are obtained. Then the researcher can use the sum of those two indices for the total SES index. An advantage of this procedure is that the canonical correlation procedure will include the test of significance for the relation between (the groups of) explanatory and proxy variables.

4. Monte Carlo Study

4.1. *Simulation Design*

This section describes a large simulation project undertaken to examine the behavior of different PCA procedures with discrete data. The measures of performance are chosen to address the accuracy of PCA in the applications of the method in ranking households by their welfare. Model (3.1) with different distributions of the underlying welfare index ξ, various coefficients Λ, various proportions of variance explained by the first PC, and other controls, as explained below, was used. This simulation was conducted in Stata software (Hilbe, 2005; Stata Corp., 2007) using a package for polychoric correlations developed by one of the authors.[4]

[4]In Stata, one can type: `findit polychoric` and follow instructions to download and install the package.

Model (3.1) was used to generate the data. The following parameters and the settings of the simulation were used:

- Total number of indicators: from 1 to 12.
- The fraction of discrete variables: from 50 percent (1 discrete, 1 continuous) to 100 percent.
- The distribution of the underlying factor: normal; uniform; lognormal; bimodal (a mixture of two normals).
- The proportion of the variance explained: 80 percent, 65 percent, 50 percent if the total number of indicators was greater than 4; 40 percent and 30 percent if the total number of indicators was greater than 7.
- The values of Λ: all ones; one or two of the discrete variables have $\lambda_k = 3$; one or two of the continuous variables have $\lambda_k = 3$; one discrete and one continuous variables have $\lambda_k = 3$.
- The number of categories of the discrete variables: from 2 to 12.
- The threshold settings: uniform (each category 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); half observations are in the central category (high kurtosis with low skewness); half observations are in the top category; random thresholds (if $\text{Prob}[y^* < z] = F(z)$, $u_1, \ldots, u_{K-1} \sim U[0, 1]$, and $u_{(1)}, \ldots, u_{(d-1)}$ is the set of order statistics from $u_1, \ldots, u_{d-1}$, then $\alpha_k = F^{-1}(u_{(k)})$).
- The sample sizes: 100, 500, 2,000, 10,000.
- Finally, and most importantly for the objective of the paper, the analyses performed: PCA on the ordinal categorical variables; PCA on the dummy variables corresponding to the individual categories, as in Filmer and Pritchett (2001); PCA on the ordinal variables with the number of the category replaced by the group means given by (3.8); PCA of the polychoric correlation matrix; PCA on the original continuous variables $x_1^*, \ldots, x_p^*$ as the benchmark (cannot be performed in the field applications).

A non-proportional random sample of all possible combinations was taken. The probability of selecting a particular combination of the simulation parameters was

$$\text{Prob}[\text{select}|\text{simulation settings}] = \exp(-(3 + 0.25p_d + 0.03p_c)) \tag{4.1}$$

where p_d is the number of discrete variables, and p_c is the number of continuous variables. An increase in the number of variables leads to the increase in computational time, both due to increased number of the polychoric ($p_d(p_d - 1)/2$) and polyserial ($p_c p_d$) correlations to be computed, and due to increase in the number of simulation settings for each extra discrete variable. This sampling procedure resulted in approximately a 1 percent sample of all settings combinations, with the total sample size of 947,434 observations, and the sum of weights (the estimate of the total population size) of 99.744 million. (This would be the total sample size should we run the simulation for each combination of parameters.) Those observations came from 189,756 unique samples (combinations of settings). Some observations were lost due to the difficulties with the numeric likelihood maximi-

zation in polychoric correlation estimation. The error messages mainly had to deal with with flat likelihoods, and also with the correlation matrix not being positive definite. Fifty-five variables were describing the settings and the outcomes. The resulting Stata file size is ≈300 Mbytes.

The primary outcome variables we consider are the internally and externally defined goodness of fit measures. The internally defined goodness of fit is what the researcher has at her disposal upon running the PCA, and the most typical measure is the proportion of the total variance explained by the first principal component (PC). The external measures of performance are those relating the estimated first PC with "the truth," i.e. ξ. The first of our measures is the Spearman rank correlation measuring agreement between the rankings of individuals produced by two variables (Conover, 1998; Hollander and Wolfe, 1999); here, the true score ξ and the PCA score. Two other measures based on the quintile groups of the theoretical and observed welfare scores are overall quintile misclassification rate and the misclassification in the first quintile, i.e. the share of observations that originally belonged to the first quintile of ξ, but were classified elsewhere by the empirical welfare measure.

For the welfare measure to be accurate, it should yield a ranking similar to the original one induced by ξ, so that the two measures rank individuals (households) in the same way. This would be reflected in high rank correlation of the empirical score with ξ, as well as in low misclassification rates. As for the explained proportion, it is usually desired to be as high as possible, but in our application, when we do know "the truth," we want it to match "the true" explained proportion as closely as possible.

4.2. *Graphical Representation*

Let us start with a graphical illustration of our findings. Here, we consider a relatively small subset of simulated data with 8 discrete and 0 continuous variables (12,880 observations out of almost 1 million total). The results for lognormal distribution of the underlying wealth variable are omitted, as they produce quite different patterns, as shown later in Section 4.3. We present several cross sections of this subset: distributional comparisons for fixed theoretical share of explained variance; performance as a function of the theoretical share of explained variance; and performance as a function of the number of categories.

Figure 1 shows the box-and-whisker plots[5] of the four performance indicator discussed in Section 4.1. The theoretical share of explained variance is fixed at 0.5, larger sample sizes of $n = 2{,}000$ and 10,000 are used, and results are pooled across different numbers of categories and threshold structures. The best performance is demonstrated by (infeasible) analysis of the original y^* variables. In the worst 25 percent of cases, the Filmer–Pritchett procedure has higher misclassification rates, either overall or in the first quintile, than the largest number any other method produced. The other three discrete methods show practically indistinguishable performance, with ordinal PCA giving slightly larger variability, as evidenced by

[5]The central line of the plot shows the median of the data. The boundaries of the box are the lower and upper quartiles. The length of each whisker is three times the distance between the median and the corresponding quartile, which leaves about 0.7 percent of the normal distribution outside the whiskers.

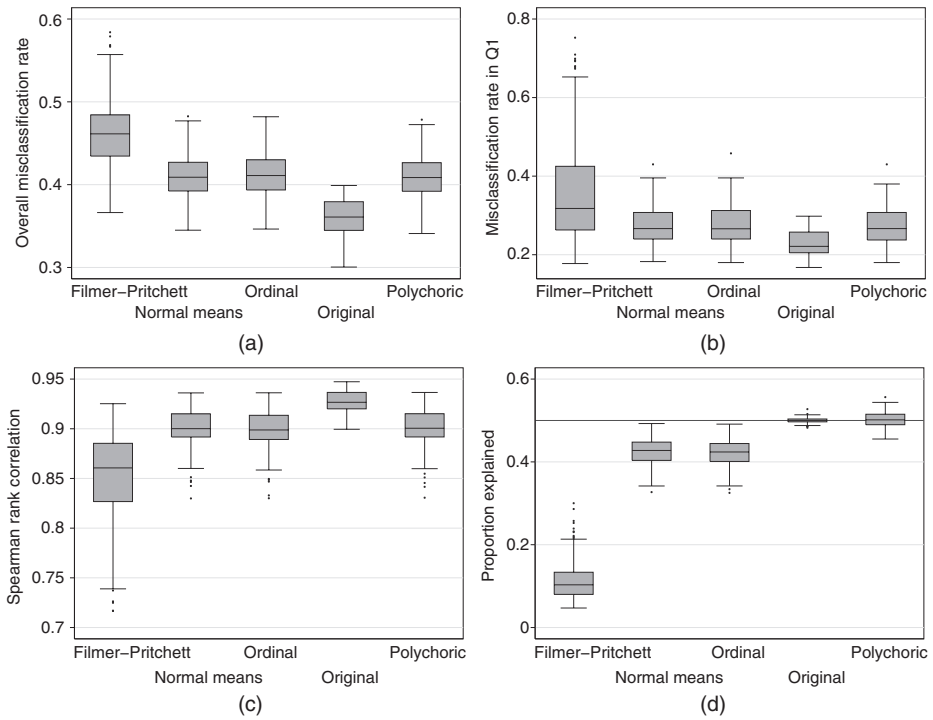

Figure 1. Box Plots for Different PCA Methods. (a) Overall misclassification rate. (b) Misclassification rate in the first quintile. (c) Spearman's ρ between the theoretical and empirical welfare measures. (d) Share of explained variance. Restrictions: 8 discrete variables, no continuous variables, sample sizes 2,000 or 10,000, lognormal distribution excluded, theoretical share of explained variance is 0.5

the size of the box. Only the analysis of the original variables and the polychoric PCA show consistency of the reported explained proportion. Other methods are demonstrating the lack of explained variance in the first PC, and the Filmer–Pritchett procedure shows particularly high bias, with no observations higher than 0.3 even though the target explained variance is 0.5.

Figure 2 shows the relation of our performance measures to the underlying theoretical proportion of the explained variance, which is the inverse of the measurement error variance. The misclassification rates (panels (a) and (b)) show almost linear decline for all methods other than Filmer–Pritchett. The latter surprisingly shows increase in variability over the whole range of the explained variances for Q1 misclassification. The reported share of explained variance (panel (d)), although consistent for the original PCA and the polychoric PCA, is underestimated by the ordinal or group means PCA, and severely biased downwards by the Filmer–Pritchett procedure. The rank correlation with the underlying welfare (panel (c)) does go up with the underlying proportion of explained variance for all methods, although the distribution of the correlations for the Filmer–Pritchett procedure demonstrates quite an extended lower tail of the distribution.

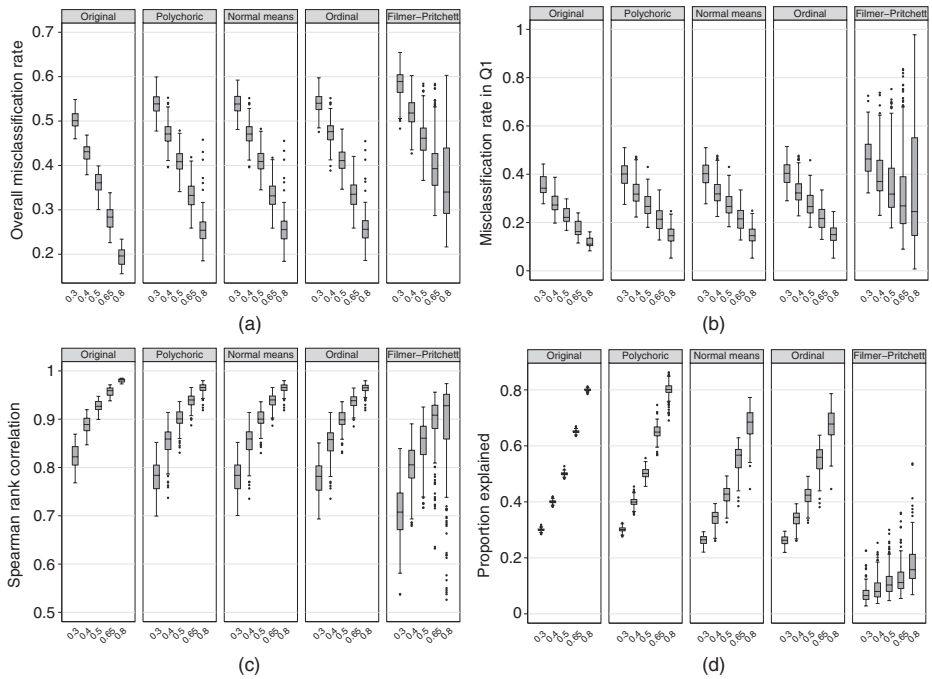

Figure 2. Relation of the Performance Measures to the Underlying Proportion of Explained Variance. (a) Overall misclassification rate. (b) Misclassification rate in the first quintile. (c) Spearman's ρ between the theoretical and empirical welfare measures. (d) Share of explained variance. Restrictions: 8 discrete variables, no continuous variables, sample sizes 2,000 or 10,000, lognormal distribution excluded. Jitter added to show structure

The next set of findings is related to the number of categories of the discrete variables used in PCA. Those are depicted in Figure 3. Note that when a variable has just two categories (e.g. ownership of an asset), the Filmer–Pritchett and ordinal PCA coincide. But as extra categories are added, the performances of the methods do differ notably. For the methods other than the Filmer–Pritchett, the four measures approach their "continuous case" limits and come to saturation at about 6 categories (except for the proportion of explained variance), consistent with findings from the quantitative sociology literature (Dolan, 1994). The performance of the Filmer–Pritchett procedure also improves with a larger number of categories, but does not get as far as in other methods until there are as many as 8 categories per variable, on average.

The most striking result is the performance of the Filmer–Pritchett procedure in terms of the reported explained variance. It declines steadily as the number of categories is increased. As the number of categories increases, more dummy variables are created by the Filmer–Pritchett procedure, thus increasing the total variation (Mardia *et al.*, 1980) of the covariance matrix. The amount of information that can be explained by the first principal component stays about the same, while the number of variables increases. The former serves as a numerator of the explained variance, and the latter as its denominator. Thus the resulting reported

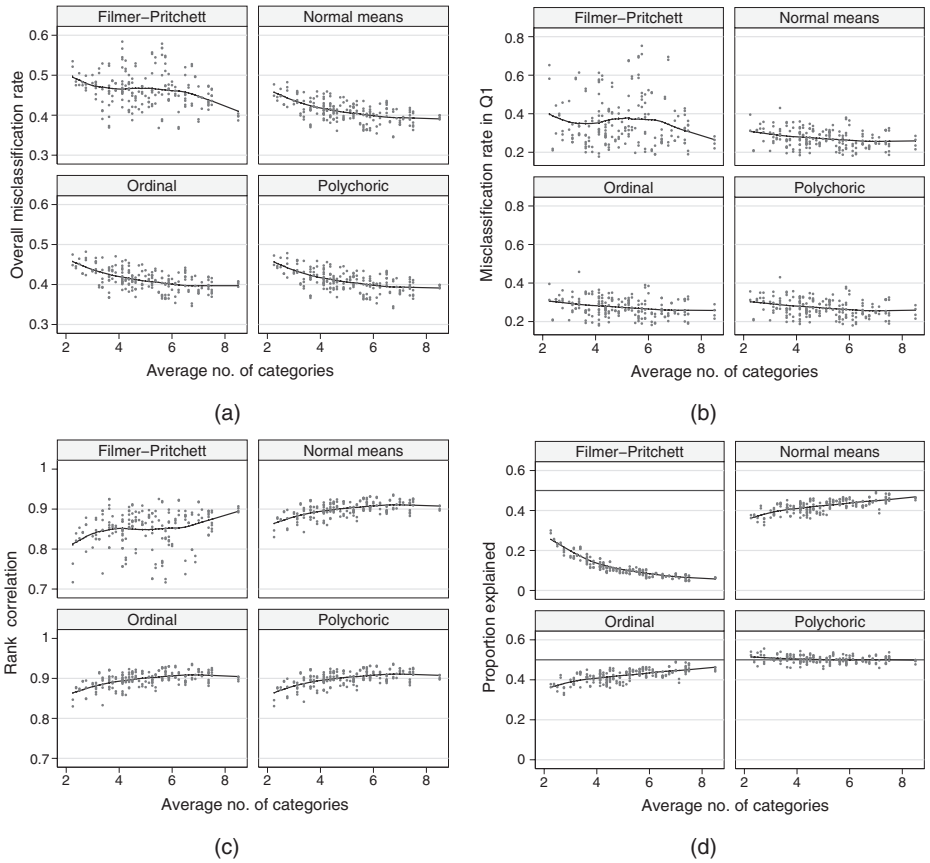

Figure 3. Relation of the Performance Measures to the Number of Categories of Discrete Variables. (a) Overall misclassification rate. (b) Misclassification rate in the first quintile. (c) Spearman's ρ between the theoretical and empirical welfare measures. (d) Share of explained variance. Restrictions: 8 discrete variables, no continuous variables, sample sizes 2,000 or 10,000, lognormal distribution excluded, theoretical share of explained variance is 0.5

share of explained variance is approximately hyperbolic in the number of categories, as observed in panel (d). As for other discrete PCA methods, the share of explained variance reported by the polychoric PCA stays on target for any number of categories, while the ordinal and the group means methods underestimate it, although improving with more categories.

4.3. *Numeric Results*

This section describes the quantitative analysis of the simulation results. Following suggestions of Skrondal (2000), we specify the regression/generalized linear model of the main effects of the simulation settings described in Section 4.1. As all the outcome measures are within a bounded range [0,1], heteroskedasticity

and non-linearity problems are pertinent in the analysis of raw data. However, most of those problems are rectified if an inverse probit transformation is used to bring the scale to $(-\infty, +\infty)$.

Table 1 presents the results of regression analysis of the complete simulated data set. The first column shows the number of observations for which a particular value of the explanatory variable is observed. The variation in the number of available observations of the analysis type is due to computational failures either with the Filmer–Pritchett or with polychoric procedure (non-positive definite matrices or lack of convergence, respectively), and of all other variables, due to randomness of the Monte Carlo procedure. Other columns of the table represent the four measures used to gauge the performance of the PCA procedures.

The rows of the table represent the most important simulation parameters and analysis types. Some other settings controlled for in the simulation summary model but not reported are the threshold structure, the factor loadings λ, and individual combinations of discrete and continuous variables. Probability weights given by the inverse of selection probabilities (4.1) were used, and covariance matrix of the estimates was corrected for clustering on the same Monte Carlo sample. The resulting standard errors do not exceed $3 \cdot 10^{-3}$ due to huge sample sizes of almost 1 million total observations, and all the results are "significant" at conventional levels.

Let us list the findings by rows of the table. High R^2s in the first row of numbers evidence that relatively few variables (64) were able to explain almost all of the variation in performance of the PCA procedures, so the remaining findings are highly reliable.

The theoretical fraction of explained proportion is the strongest factor, with greater values associated with better performance.

The next block of the table compares the four versions of PCA to the (unfeasible in the field) analysis of the underlying continuous variables as the benchmark. In all four measures, the Filmer–Pritchett procedure shows the greatest difference from that benchmark. The marginal effect at the "average" setting for misclassification rates is about 30 percent, compared to 18 percent for other discrete methods. The latter three show very similar performance except for the share of explained variance, where the coefficient of the polychoric method is lower, which indicates smaller biases in estimating the explained variance.

The importance of the next block, the distribution of the underlying factor, is in showing that heavy tailed distributions of the true wealth index affect negatively the PCA procedures. The coefficients for lognormal distribution are always in the direction of performance deterioration. The marginal effect is about 15 percent for overall misclassification rate, and 30 percent for misclassification in the first quintile. All other distributions have rather mild effects, including the skewed bimodal distribution that, however, produced some difficulties for classification in the first quintile.

The effect of the number of categories is mostly pronounced for the Filmer–Pritchett procedure, where a greater number of categories reduces the reported share of explained variance, as shown graphically and discussed in the previous section. The effects of increasing the number of categories with other methods are positive.

TABLE 1

MONTE CARLO SIMULATION RESULTS: PERFORMANCE OF DIFFERENT VERSIONS OF PCA ON DISCRETE DATA

	No. Obs.	Share of Explained Variance	Rank Correlation with ξ	Overall Misclassification Rate	Misclassification in Q1
R-squared		0.93	0.93	0.90	0.81
Explained variance		–0.623	1.927	–1.403	–1.269
Analysis type					
Original: base	189,756	0	0	0	0
Filmer–Pritchett	189,511	–0.557	–0.333	0.237	0.243
Group means	189,328	–0.339	–0.226	0.172	0.144
Ordinal	189,511	–0.345	–0.231	0.176	0.147
Polychoric	189,328	–0.158	–0.223	0.170	0.142
Distribution of ξ					
Normal: base	233,688	0	0	0	0
Lognormal	237,222	–0.203	–0.735	0.422	0.837
Bimodal	234,612	0.005	–0.097	0.036	0.209
Uniform	241,912	0.024	0.065	–0.051	0.047
Average no. categories					
Filmer–Pritchett	189,511	–0.135	0.001	–0.004	0.008
Ordinal, polychoric, group means	568,167	0.015	0.020	–0.015	–0.011
Log samplesize		–0.001	0.010	–0.006	–0.010
Observed variables					
Largest difference		⟨1, 1⟩–⟨10, 0⟩ 0.319	⟨2, 0⟩–⟨8, 4⟩ –1.368	⟨2, 0⟩–⟨8, 4⟩ 0.898	⟨1, 1⟩–⟨8, 4⟩ 0.790
⟨4, 2⟩ vs. ⟨4, 3⟩		–0.021	–0.102	0.074	0.070
⟨4, 2⟩ vs. ⟨5, 2⟩		0.041	–0.058	0.042	0.037
⟨8, 0⟩ vs. ⟨8, 1⟩		–0.018	–0.103	0.067	0.066
⟨8, 0⟩ vs. ⟨9, 0⟩		0.005	–0.067	0.044	0.045

Notes: Other controls include: the threshold structure; the factor loadings; the number of discrete and continuous variables. Total number of observations is 947,434. Number of clusters is 189,756. Cluster corrected standard errors are less than $3 \cdot 10^{-3}$ for all reported coefficients; clusters are defined by the same sample used in different versions of PCA. The notation ⟨8, 4⟩ means a model with 8 discrete and 4 continuous variables, etc.

Interestingly, the sample size does not have much effect on the results. The marginal effect of increasing the sample size from the smallest setting of 100 to the largest setting of 10,000 produces the main effect of the improvement in classification rates of about 2 percent.

The last block of the table shows the effect of the number and the type of variables. The notation $\langle p_d, p_c \rangle$ stands for data configuration with p_d discrete and p_c continuous variables. The largest difference across the estimated coefficients is usually between a model with two indicators, and a model with 12 indicators. Thus the differences between the misclassification rates for different number of variables may be as large as 64 percent vs. 26 percent for the overall rate, and 50 percent vs. 17 percent for the first quintile. The next few rows show marginal effects of adding a continuous or a discrete variable. The improvement due to a continuous variable is larger than that for a discrete one by some 60–80 percent. This can be viewed as a crude measure of the information losses due to discreteness: roughly speaking, 10 discrete variables contain about as much information, for the PCA purposes, as 6 continuous ones do.

5. Sensitivity Analysis

Another simulation was performed to study robustness of the ordinal procedures to misspecifications of the ordinal structure (that is, incorrect ordering of categories). In this simulation, we shuffled some of the ordinal categories, which would represent an incorrect ordering of the categories by the researcher. For instance, while a tin roof is undoubtedly superior to a clay or straw roof, the comparison between the latter two may not be very obvious, and may in fact go either way depending on the local climate and availability of materials. The problem is not likely to arise with binary indicators representing ownership of a certain durable good.

In this simulation, we used a subset of the primary settings of the main simulation. The distribution of the underlying index variable was normal; the threshold structure was uniform; and the loadings structure was uniform, as well. The greatest problems would be caused in the situation where there is little additional information available to recover the appropriate rankings, thus we focused on situations with few variables and few categories per variable. This was expected to be beneficial for the Filmer–Pritchett procedure, which does not make any assumptions about the order of categories.

For each combination of settings, 1,200 Monte Carlo replications were simulated, and six analyses performed on each one: the PCA of the original continuous variables (not feasible to the researcher, and to be used as the baseline for comparison); ordinal and polychoric PCA of the original unswapped ordinal variables (representing the situation where the categories were classified properly); ordinal and polychoric PCA of the permuted data (representing the situation with misclassified categories); and the Filmer–Pritchett PCA on dummy variables that ignores the ordinal properties, and thus identical for the original and permuted data. We studied six misclassification schemes. Here, we report the results for the schemes that led to the greatest and the smallest differences between correctly and incorrectly specified PCA models.

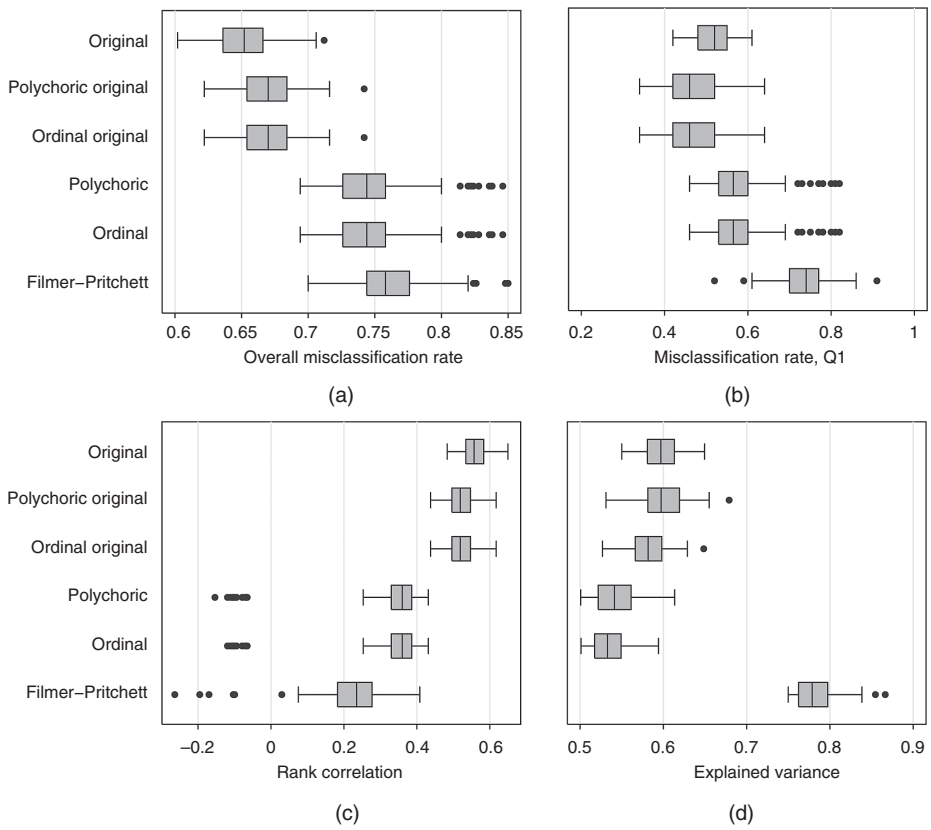

Figure 4. Performance of Various PCA Procedures. Sample size is 500, the theoretical share of explained variance is 0.6, two variables out of four have misclassified ordered categories

Figure 4 shows the relative performance of different PCA methods for the following misclassification scheme. Four variables are used. X_1 is ordinal with 4 categories, X_2 is ordinal with 3 categories, X_3 and X_4 are binary. Categories 2 and 3 of X_1 and X_2 are swapped, which are the middle two categories of X_1, and the middle and the top categories of X_2. Sample size is $n = 500$, the population proportion of explained variance is 0.6. The first three boxplots in each panel are infeasible analyses, and given as the baseline. The last three are feasible polychoric, ordinal, and the Filmer–Pritchett versions of PCA. Clearly, the Filmer–Prithcett procedure does not perform better despite the biases to the ordinal procedures introduced by misclassification. While overall misclassification rates are comparable among the three feasible procedures, the misclassification rate in the first quintile and rank correlation with the true wealth ranking are notably worse, the latter also being affected by a group of outliers with negative rank correlation for the ordinal methods. Moreover, the Filmer–Pritchett procedure boasts a fraction of explained variance that is significantly biased *upwards*, which is an unusual outcome in light of the earlier findings. While the methods based on unperturbed data produce overall misclassification rates of about 65–67 percent and the first

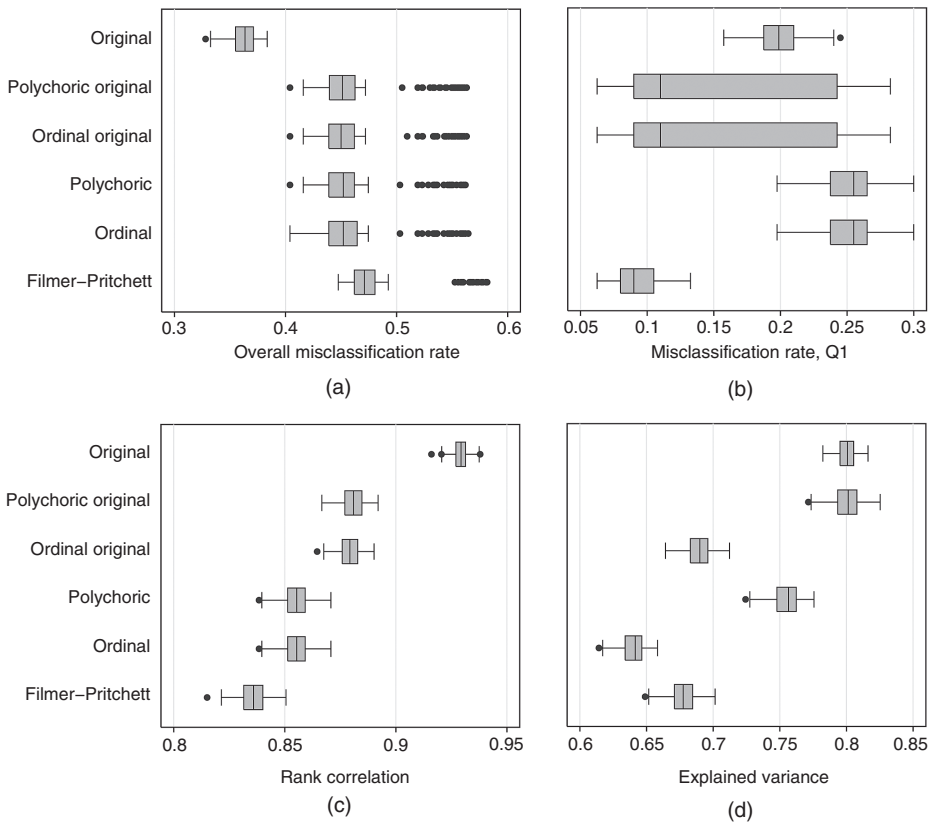

Figure 5. Performance of Various PCA Procedures. Sample size is 2,000, the theoretical share of explained variance is 0.8, one variable out of three has middle categories misclassified

quintile misclassification rates of about 45–50 percent (with the PCA based on the original continuous variables surprisingly performing worse on average than the ordinal methods), the feasible methods suffer a deterioration of at least 10 percent in those rates. As expected, the polychoric and ordinal methods produce very close results. The advantage of the polychoric methods in producing more accurate share of explained variance results dissipates.

The mildest misclassification effects were obtained in the following scheme. Three variables were created. X_1 is ordinal with 4 categories, the middle two categories of X_1 are swapped; X_2 and X_3 are binary. The results for sample size of $n = 2{,}000$ and the population proportion of explained variance of 0.8 are shown in Figure 5. The Filmer–Pritchett procedure handled the low end of the distribution better than other methods, but overall rankings were reproduced better by polychoric and ordinal analyses. The polychoric procedure reported a higher proportion of explained variance, which was unbiased for the original unshuffled data, but still low for the misclassified data. The ordinal and the Filmer–Pritchett PCA underestimate that proportion.

Among the other four misclassification schemes, there was one scheme where the Filmer–Pritchett produced universally better results on all measures (three variables: X_1 is ordinal with 4 categories, categories 2 (lower middle) and 4 (top) are swapped; X_2 and X_3 are binary), and one scheme where ordinal and polychoric were producing identical results universally better than those from the Filmer–Pritchett procedure (two variables: X_1 is ordinal with 4 categories, with categories 2 (lower middle) and 4 (the top) swapped, and X_2 is ordinal with three categories, none of those changed). The other two settings were producing mixed results. Overall, the conjecture that the ordering-free nature of the Filmer–Pritchett procedure would produce better results when the ordering of categories is not specified properly was not confirmed.

6. Empirical Illustration

In this section, we demonstrate the different approaches with the example of Bangladesh DHS 2000 data. The sample design is a stratified clustered two-stage procedure with varying probabilities of selection. The household data set contains 9,821 observations in 341 communities. The variables of interest are presented in the first column of Table 2. There are 5 binary ownership variables and 6 ordinal variables containing information on the housing quality. Some of the less populated categories that were viewed to represent the similar levels of quality were collapsed into a single one. The sources of drinking and non-drinking water were converted from 6 categories down to 4; the type of toilet facility, from 5 to 4; and the main wall material, from 4 to 3. The different versions of the PCA were performed on the same data set with unified categories. While Stata does not allow using probability weights in `pca` command, using frequency weights `[fw = weight variable]` still produces valid estimates of the covariance matrix. The polychoric package, on the other hand, does support the appropriate probability weights, which can be specified by `[pw = weight variable]` statement.

6.1. *The Filmer–Pritchett Procedure*

The PCA based on the binary category indicators reported 24 percent of the total variance to be explained by the first PC. The factor loadings (see Table 2) show the desirable monotonicity, but one should keep in mind that they are relative to the base category that is getting the implicit weight of zero, and in this case, the base category for all of the variables was the lowest category labeled 1. The scree plot shows that the first component is highly significant, and at least three other components might contain some information about the correlation structure of the original set of variables. In this and all other PC analyses, the scores of the first principal component were shifted so that the lowest score is set to zero. The distribution is highly asymmetric and leptokurtic. The Gini concentration coefficient is 0.58, which puts this distribution into very high welfare inequality range. According to the WIID database (WIDER, 2000), the highest Gini index across a range of countries is 0.59 (South Africa in 1995), while the lowest is 0.25 (Sweden in 2000). These are the Gini indices of income rather than welfare distribution, however. The WIID entries for Bangladesh suggest income

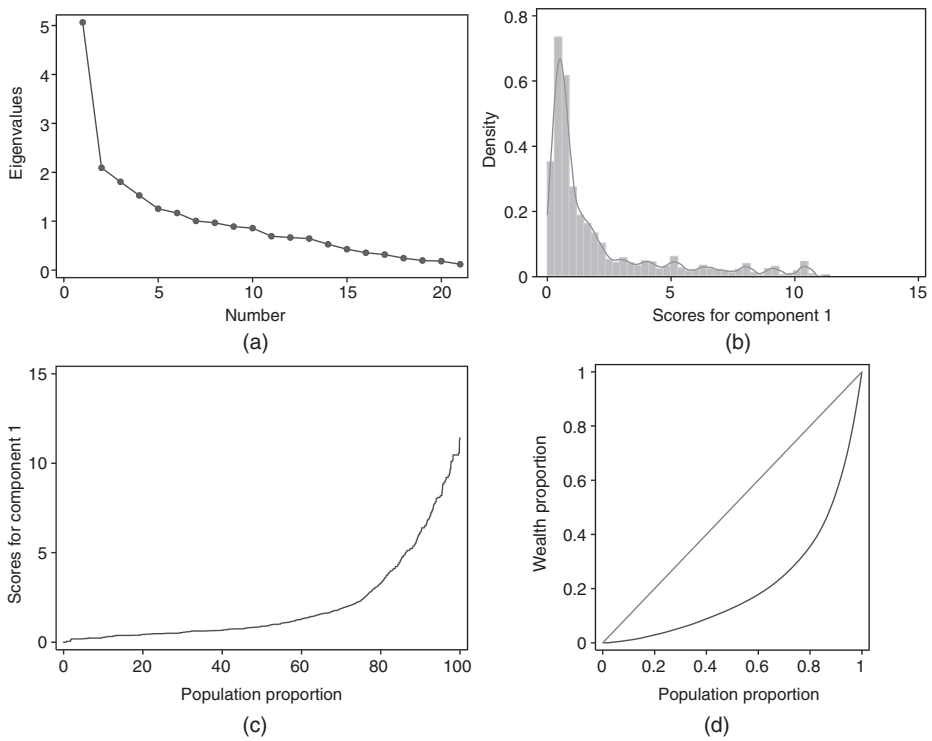

Figure 6. Diagnostic Graphs for the Filmer–Pritchett PCA on Bangladesh DHS 2000 Data. (a) Scree plot: the first component is significant, another three may also contain non-trivial information. (b) Distribution of the first PC. Skewness = 1.86, kurtosis = 5.61. (c) Rank plot of the first PC. (d) Lorenz curve of the first PC. Gini = 0.5818

inequality of 0.336 (WDI database) to 0.428 (Bangladesh Bureau of Statistics). See http://www.wider.unu.edu/wiid/data/BGD.htm.

6.2. *Ordinal PCA*

The PCA based on the ordinal variables (recoded to start at 1 and have steps of 1) showed 39 percent of variance to be explained by the first component. All indicators have about the same factor loadings of about 0.3–0.4, except for the bicycle and motorcycle indicators that have weights of about 0.1. Based on the scree plot, the components after the second one are indistinguishable from noise. The distribution of the first PC shows considerable asymmetry. The Gini concentration coefficient for this distribution is 0.35, which puts it into the moderate inequality range, and is largely consistent with the WIID figures.

6.3. *Polychoric PCA*

Polychoric PCA with 11 variables, 55 correlations to be estimated and almost 10,000 observations took about 25 minutes on a 1.6 GHz Windows XP computer. It produced what we believe to be the most accurate 56 percent of the variance

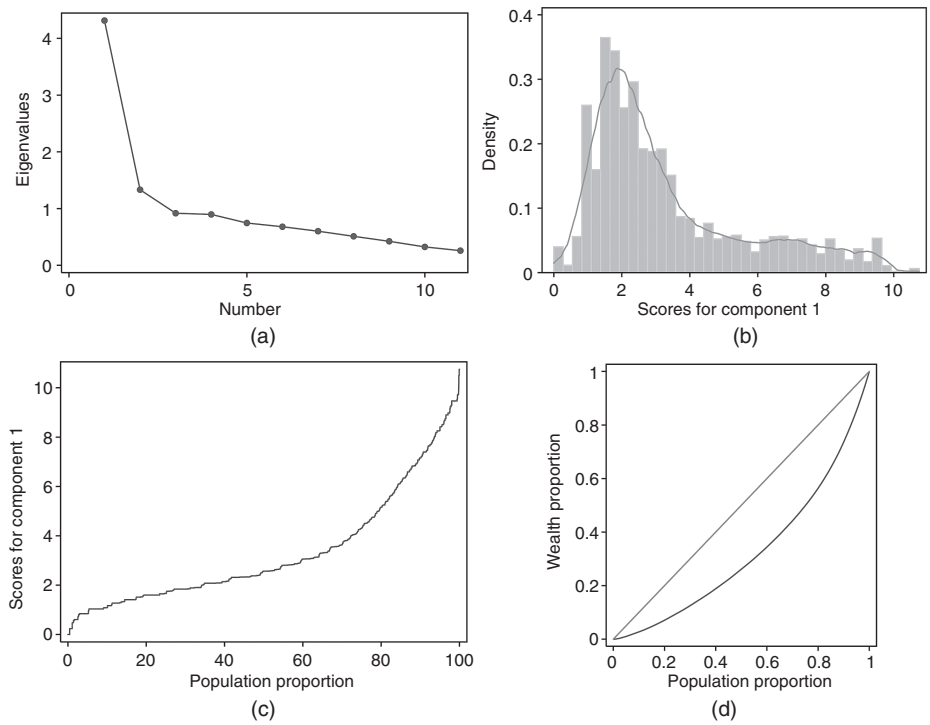

Figure 7. Diagnostic Graphs for the Ordinal PCA on Bangladesh DHS 2000 Data. (a) Scree plot: the first component is significant, the second component is marginally significant. (b) Distribution of the first PC. Skewness = 1.16, kurtosis = 3.45. (c) Rank plot of the first PC. (d) Lorenz curve of the first PC. Gini = 0.3536

explained by the first component. The (normalized) factor loading are in 0.25–0.40 range with the exception of the bicycle variable with a loading of 0.13. The scree plot shows that the first component is highly significant, and the second component might contain some non-noise information, too. The skewness and kurtosis are slightly less than for the ordinal PCA, and the Gini coefficient is somewhat milder at 0.31, and lower than the reported inequality figures in WIID.

6.4. *Comparisons*

The comparison of the welfare index weights implied by the three different procedures is given in Table 2. The weights in columns 2 and 3 are the weights for standardized variables, i.e. the ones based on the analysis of the polychoric correlations matrix. In actual applications, one would need to rescale the variables so that they have a variance of 1 before the weights from the Filmer–Pritchett or ordinal PCA column can be applied.[6] The coefficients from the column labeled "Ordinal PCA" are on the same scale as the coefficients in the "Eigenvector"

[6]In Stata, this can be easily achieved by `egen . . . = std ( . . . )` command. In a common case when the principal component analysis is to be performed and scores are to be computed on the same data, the latter can be achieved by the standard Stata `predict` command following immediately after `pca`.

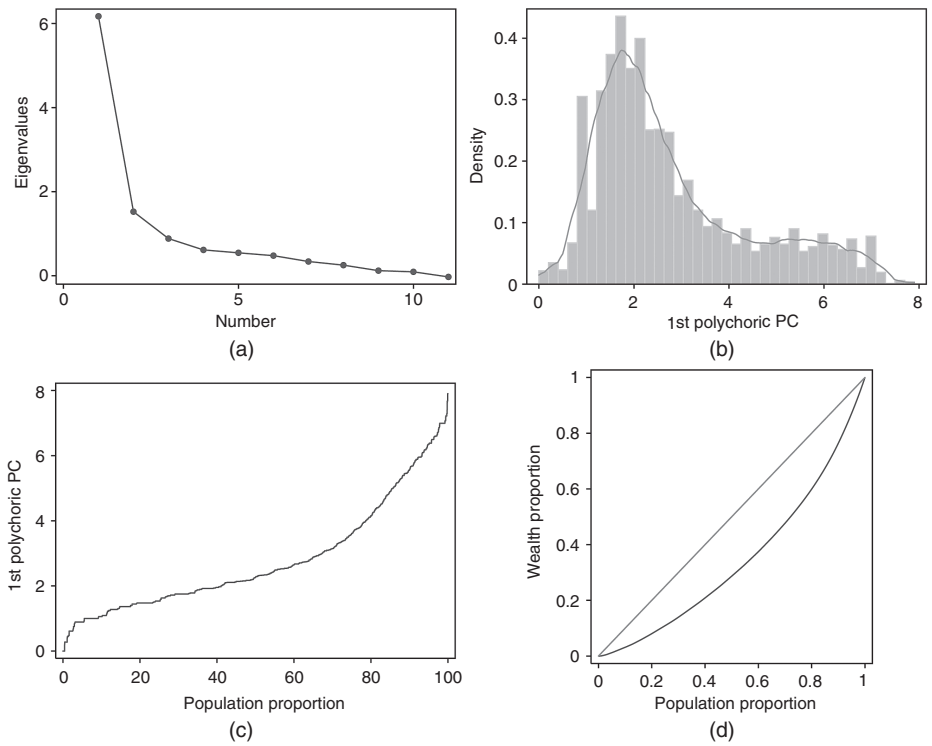

Figure 8. Diagnostic Graphs for the Polychoric PCA on Bangladesh DHS 2000 Data. (a) Scree plot: the first component is significant, the second component is marginally significant. (b) Distribution of the first PC. Skewness = 0.99, kurtosis = 3.11. (c) Rank plot of the first PC. (d) Lorenz curve of the first PC. Gini = 0.3123

subcolumn of polychoric PCA, while the coefficients from the Filmer–Pritchett analysis should show the patterns similar to the scoring weights subcolumn of polychoric PCA. In the latter subcolumn, the output of Stata's `polychoric` has been used that already accounts for the scale calibration.

Table 2 shows why the Filmer–Pritchett procedure does not necessarily produce reliable results. Of all the multinomial ordinal variables, the Filmer–Pritchett scores show the desirable monotonicity only for flooring material. For all others, the second category that is expected to be superior to the first one has a weight lower than the inferior category (negative vs. 0). Thus, the weights produced by the Filmer–Pritchett procedure are not necessarily consistent with the ordering information.

The fact that the eigenvector entries in ordinal and polychoric analyses are almost all within a range of 0.3–0.4 sheds light on why the sum of assets is performing reasonably well in Bollen *et al.* (2001), as well as in our fertility analysis in Section 6.6. The simple sum of assets seems to be highly correlated with the first principal components defined by either the ordinal or polychoric versions of PCA, but it only gives 22 distinct numeric values, compared to 1,336 by the ordinal/polychoric score.

TABLE 2

WELFARE INDEX WEIGHTS FOR DIFFERENT VERSIONS OF THE PCA

Variable	Filmer–Pritchett PCA	Ordinal PCA	Polychoric PCA		2nd Ordinal PC
			Eigenvector	Scoring Weight	
Source of drinking water		0.2919	0.2856		–0.4920
Surface well, lake, pond, stream (1)	0			–0.6267	
Tube well (2)	–0.2617			–0.0130	
Piped outside (3)	0.0859			0.4604	
Piped inside (4)	0.3150			0.5980	
Source of non-drinking water		0.3095	0.2571		–0.4137
Surface well, lake, pond, stream (1)	0			–0.3077	
Tube well (2)	–0.1277			0.0786	
Piped outside (3)	0.0858			0.3829	
Piped inside (4)	0.3421			0.5076	
Type of toilet facility		0.3094	0.2917		0.1931
No facility (1)	0			–0.4084	
Open latrine (2)	–0.0649			–0.1317	
Pit latrine (3)	–0.0752			0.0371	
Water sealed (4)	0.0044			0.2228	
Septic tank/toilet (5)	0.3089			0.5104	
Has electricity	0.2837	0.3506	0.3451	0.5671	0.0212
Has radio	0.1640	0.2272	0.2443	0.4019	0.3791
Has television	0.3016	0.3584	0.3663	0.6541	0.1240
Has bicycle	0.0441	0.1011	0.1278	0.2231	0.5362
Has motorcycle	0.1116	0.1365	0.2728	0.6838	0.3012
Main floor material		0.3986	0.3918		–0.0996
Earth/bamboo (1)	0			–0.1120	
Wood (2)	0.0051			0.3969	
Cement/concrete (3)	0.3718			0.6042	
Main wall material		0.3773	0.3417		0.0191
Natural (1)	0			–0.2112	
Rudimentary/tin (2)	–0.0754			0.2070	
Brick/cement (3)	0.3532			0.5097	
Main roof material		0.3004	0.3054		0.0372
Earth/bamboo (1)	0			–0.4227	
Wood (2)	–0.1130			0.0530	
Cement/concrete (3)	0.2909			0.5509	
% explained variance	24.11%	39.23%	56.09%		12.11%

Another direction of comparisons is the relative classification of the data points into quintiles. The performance of different methods relative to the polychoric score is given in Table 3. In the second half of the table, quintiles are computed taking into account the sampling weights. The ordinal score and the simple sum are not particularly different from the polychoric score, but the Filmer–Pritchett score has an overall misclassification rate relative to the polychoric score of 45 percent for unweighted and 59 percent for weighted data. The 494 observations common for the first unweighted quintiles of both the polychoric and the Filmer–Pritchett scores are generated by only three unique combinations of the underlying indicators, with 383 observations being the largest cluster of identical observations across the whole data set.[7] For both the polychoric and the Filmer–Pritchett scores, the overlapping observations of the first quintile are actually close to the boundary of the second quintile. If the data are weighted according to the sampling weights, then the Filmer–Pritchett and polychoric scores do not commonly classify any single observation in the poorest quintile! The scatterplot of the two scores in Figure 9 sheds some light on the picture. (The size of the bubble is proportional to the sum of weights for the corresponding combination of the variables.) It confirms that at high SES/welfare levels, the two procedures do show agreement. At lower levels, the two procedures differ markedly, primarily because of the lack of variability in the data. Many households have the same observed characteristics, such as mentioned in footnote 7 but due to different weights given by different procedures, and especially due to ordering inconsistent weights given by the Filmer–Pritchett procedure, the rankings differ quite markedly.

6.5. *The Second Principal Component*

The last column of Table 2 reports the factor loadings for the second principal component produced by the ordinal PCA. While the first PC usually describes some measure of "size," in this case, welfare, the second and further components, if they are at all informative, describe the "structure," i.e. in what ways the households ranked similarly by the first component differ most markedly. By geometry and algebra of PCA, those next components would be more heavily loaded by the variables whose weights were relatively low in the first component. That is, by and large, the picture here, as well. Most variables that had high loadings on the first PC (electricity, TV, dwelling materials) have low loadings on the second component, although the sources of (both drinking and non-drinking) water variables contribute substantially to the second component, too. The transportation means variables that had the lowest loadings on the first component show much higher loadings on the second component. For households of comparable SES/welfare levels, the primary differences are due to access to clean water and transportation needs. The households with high levels of the second principal components would have a bicycle and/or a motorcycle, but poor water sources. This hints that the second component may have a meaning of the overall

[7]These households use tube well for drinking and non-drinking water; they do not have toilet facilities, electricity, radio, TV, bicycle, or motorcycle; the floors are made of bamboo or wood; the walls are made of natural materials; and the roof is made of wood.

TABLE 3

RELATIVE MISCLASSIFICATION RATES VS. THE POLYCHORIC SCORE

Quintiles of Polychoric Score	Quintiles of Unweighted Data						Quintiles of Weighted Data					
	1	2	3	4	5	Total	1	2	3	4	5	Total
. . . vs. simple sum quantiles												
1	2,080	60	0	0	0	2,140	1,686	454	0	0	0	2,140
2	198	1,591	552	0	0	2,341	186	1,201	929	25	0	2,341
3	0	42	1,356	572	0	1,970	0	22	768	288	0	1,078
4	0	0	16	1,387	217	1,620	0	0	119	1,622	190	1,931
5	0	0	0	15	1,735	1,750	0	0	0	30	2,301	2,331
. . . vs. ordinal score quantiles												
1	2,272	27	0	0	0	2,299	1,858	0	0	0	0	1,858
2	6	1,612	46	0	0	1,664	14	1,552	286	0	0	1,852
3	0	54	1,847	48	0	1,949	0	125	1,519	16	0	1,660
4	0	0	31	1,894	35	1,960	0	0	11	1,924	21	1,956
5	0	0	0	32	1,917	1,949	0	0	0	25	2,470	2,495
. . . vs. Filmer–Pritchett score quantiles												
1	494	855	645	46	0	2,040	0	893	652	138	0	1,683
2	927	450	395	126	0	1,898	933	262	536	131	0	1,862
3	579	302	563	522	5	1,966	619	366	232	541	1	1,759
4	278	86	316	1,124	154	1,958	311	142	372	920	228	1,973
5	0	0	5	156	1,798	1,959	9	14	24	235	2,262	2,544
Total	2,278	1,693	1,924	1,974	1,952	9,821	1,872	1,677	1,816	1,965	2,491	9,821

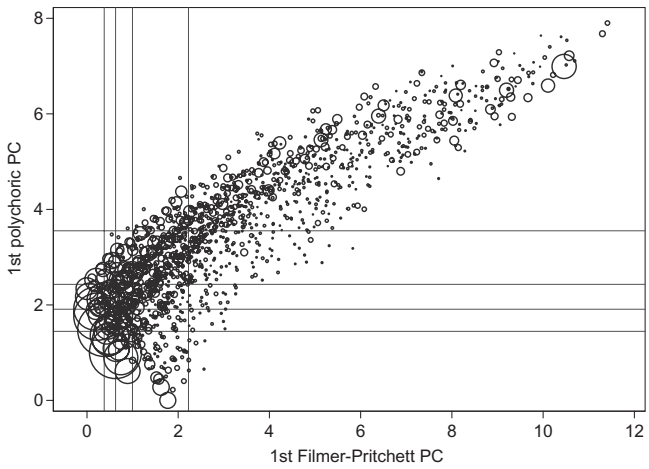

Figure 9. Scatterplot of the two scores. The area of a bubble is proportional to the sum of sampling weights. The lines represent the 20, 40, 60 and 80th percentiles of the corresponding scores

development, or urbanization, with more urbanized areas having better public transportation and higher density infrastructure, as well as centralized water and sewage systems, leading to low values of the second principal component.

6.6. *External Validity*

Suppose a researcher puts forward a social theory predicting that a certain unobserved concept, such as SES, has a causal effect on observed phenomena, such as health behaviors. For instance, it might be expected that fertility decreases with SES due to higher opportunity costs and bargaining powers of women in households with higher SES. Both sides of this relation can be measured: the SES, in any of the ways described in this paper and elsewhere; and fertility, by the number of births in the recent past (DHS uses 1, 3 and 5 year frames), or the total number of children ever born. If the indicator is exhibiting the empirical relation in accordance with theory, then it is said to possess *external validity* (Hand, 2004).

The PCA-based methods of assigning weights for the SES index only satisfy a weaker requirement of *internal validity*. The latter means that the social theory predicts that the variables will be varying together. The reported proportion of explained variance is a common measure of internal validity, the higher the better.

Table 4 explores the issue by looking at the probability of giving birth in the last three years. This is the same measure as used by Bollen *et al.* (2006) in their comparisons. This event was recorded for 36 percent of women in the sample.

The three methods used above are compared, as well as the simple sum of assets. The latter makes use of occasional values higher than 1 in the data (e.g. for the number of TV sets or bicycles in the household). Also, the ordinal variables were scaled to be 1, 2, 3, . . . in computing the sum of assets index. All SES scores were standardized to have the mean 0 and variance 1, to simplify comparisons.

TABLE 4

PERFORMANCE OF VARIOUS SES INDICES IN EXPLAINING FERTILITY IN WOMEN OF REPRODUCTIVE AGE, BANGLADESH 2000

Regressor	Sample %	Filmer–Pritchett	Ordinal	Polychoric	Sum of Assets
SES index		–0.115	–0.169	–0.169	–0.167
		(0.029)	(0.031)	(0.031)	(0.031)
z-ratio		–3.93	–5.40	–5.45	–5.34
Marginal effect at base, %		–4.6%	–6.7%	–6.7%	–6.7%
Controls					
Christian	11.2%	–0.140	–0.154	–0.157	–0.156
		(0.058)	(0.058)	(0.058)	(0.058)
Primary education	28.5%	–0.219	–0.190	–0.187	–0.184
		(0.043)	(0.043)	(0.043)	(0.043)
Secondary education	21.4%	–0.247	–0.175	–0.170	–0.169
		(0.051)	(0.053)	(0.053)	(0.054)
Higher education	4.2%	–0.037	0.060	0.063	0.062
		(0.104)	(0.107)	(0.107)	(0.106)
Urban	6.3%	–0.032	0.023	0.008	0.009
		(0.087)	(0.084)	(0.082)	(0.082)
Town	13.8%	–0.038	0.013	0.014	0.008
		(0.066)	(0.065)	(0.065)	(0.065)
Age 10–19	16.1%	0.290	0.274	0.273	0.274
		(0.056)	(0.056)	(0.056)	(0.056)
Age 20–24	18.4%	0.406	0.401	0.400	0.401
		(0.052)	(0.052)	(0.052)	(0.052)
Age 30–34	15.4%	–0.429	–0.422	–0.422	–0.421
		(0.051)	(0.051)	(0.051)	(0.051)
Age 35–39	12.7%	–0.933	–0.927	–0.926	–0.926
		(0.062)	(0.062)	(0.062)	(0.062)
Age 40–49	18.8%	–1.680	–1.665	–1.663	–1.662
		(0.075)	(0.075)	(0.075)	(0.075)
Intercept		0.078	0.035	0.033	0.032
		(0.052)	(0.051)	(0.051)	(0.051)

Notes: Reported entries are coefficients in the probit regression for probability of having given birth in the last three years. Sampling weights are used. The standard errors are corrected for clustering in the sample design. Base categories: no education (45.9%), rural/countryside (79.9%), age 25–29 (18.7%).

Before interpreting the results, it should be noted that in probit models, only the combinations β/σ are identified where β is the regression coefficient and σ is the standard deviation of the probit error terms. (See discussion of the issue in Appendix D.) The results for ordinal, polychoric and the simple sum indices show very good agreement, with coefficients agreeing in two decimal points. The Filmer–Pritchett procedure demonstrates a smaller coefficient, smaller significance and a smaller marginal effect than the other methods considered. It also produced somewhat different coefficients of other control variables, especially for the education variables. The significant variables have coefficients of larger magnitude than with the other three methods. In other words, the variability in the outcome had to be explained by those variables rather than by the SES proxy. This is a complementary effect to the lower impact of the (standardized) SES proxy. If we are willing to consider the stronger reported effects of SES to be evidence of better performance of the measure, as suggested by Bollen *et al.* (2006), then other methods outperform the Filmer–Pritchett procedure.

7. Conclusion

This paper was motivated by recent examples applying principal component analysis in the development economics literature, and it investigated several ways to use categorical (in particular, ordinal and binary) variables in PCA. As far as the distributions of the indicators are non-normal, some of the properties of the principal components no longer hold or need to be revised. Some complications to the principal component analysis due to the categorical nature of the variables include biases to the covariance structure, and hence the factor loadings, and smaller reported proportion of explained variance.

We discussed several options that may be useful in performing the principal component analysis in the presence of categorical variables: using ordinal variables as if they were continuous; using the group means implied by a normal distribution; using the dummy variables for categories as suggested by Filmer and Pritchett (2001); and using the polychoric correlations. We designed and conducted a large simulation study to compare the performance of different discrete PCA methods under different scenarios. Our main conclusions stemming from the analysis of the simulation data are as follows.

If there are several categories related to a single factor, such as the access of hygienic facilities or the quality of the dwelling materials, dividing the variable into a set of dummy indicators as suggested by Filmer and Pritchett (2001) leads to deterioration of performance, according to all of the performance measures we used. The explained variance is most heavily affected (underestimated), and more so with more categories of the ordinal variables. Even though the goodness of fit of the Filmer–Pritchett procedure improves as we add more variables, the method does not achieve the performance shown by the other methods. We thus believe that the researcher would be better off using the ordinal variables as inputs to PCA. If the variables do not come in a "standard" way, such as 1, 2, . . . (Likert scale) with roughly equal distances between categories, it is worth recoding them that way, so that those distances are not very different. Model-based category weights (referred to as "group means" in our analysis) show but slight improvement in performance compared to the "standard" Likert-scale ordinal coding, so "naïve" coding is strikingly robust to the arbitrary assumption of the distance between categories being 1.

The gain from using computationally intensive polychoric correlations compared to the PCA on ordinal data is only related to more accurate estimation of the proportion of explained variance that other methods tend to underestimate. The misclassification rates, as well as rank correlations of the theoretical and empirical welfare indices, are not substantially different among the ordinal, group means, and polychoric versions of PCA. Thus the use of the polychoric method is only justified if the proportion of explained variance is used for important reporting and/or decision-making purposes.

The performance of PCA also depends on a large number of factors. As expected, the most important ones are the underlying proportion of explained variance in the population, which controls the strength of relation between the welfare and its indicators, and the number of variables available to the researcher.

As they increase, the performance improves. A heavy tailed distribution of the underlying factor will likely lead to a notable degradation of the PCA performance. The goodness of fit improves as the number of categories per factor increases, although the returns are not so great once the researcher can distinguish about 5 categories in each of the variables. Other factors in the simulation design, such as the placement of thresholds, were found to be of marginal importance for performance of PCA.

Those results are by and large similar to what is known in other areas of multivariate quantitative social sciences dealing with ordinal variables. They also confirm the expectations outlined in simpler settings in the theoretical part of the paper. They should also be viewed in the light of the particular data generating model.

A sensitivity analysis was performed to study the performance of the studied procedures when the ordering of the categories is not specified correctly. The evidence was not conclusive: for some forms of misspecification, the Filmer–Pritchett procedure was producing better results; for others, the ordinal and polychoric procedures were working better; and for yet others, different outcome variables ranked the two (groups of) methods differently.

The empirical analysis using Bangladesh DHS data demonstrated quite notable differences between the scores obtained from the Filmer–Pritchett procedure, on one hand, and the ordinal, polychoric, and sum of asset indices, on the other hand. The distributions of the first principal component arising from the Filmer–Pritchett procedure showed greater skewness and kurtosis than those of other procedures, and led to higher Gini coefficients. While the ordinal, polychoric and sum of assets scores produced rankings of the households that were quite similar, the Filmer–Pritchett procedure disagreed with them, especially in the lower part of the distribution.

When the proposed SES indices were put into action of explaining fertility, the ordinal, polychoric and sum of assets scores produced nearly identical methods, while the Filmer–Pritchett score had lower significance and probably produced a model of lower explanatory power.

Our general recommendations are then as follows. If there is a reliable and well established ordering of categories, the ordinal PCA should be used. However, if the proportion of explained variance is of importance, the polychoric method should be used. These methods tend to outperform the Filmer–Pritchett procedure when the orderings are correctly specified. The Filmer–Pritchett procedure is not recommended unless there is absolutely no information about the ordering of categories.

The appendices to this paper are available at http://web.missouri.edu/~kolenikovs/papers/roiw-309-appendices.pdf. The appendices are also available at the following Wiley-Blackwell website: http://dx.doi.org/10.1111/j.1475-4991.2008.00309.x.

References

Abadir, K. M. and J. R. Magnus, *Matrix Algebra*, Cambridge University Press, New York, 2005.

Anderson, T. W., "Asymptotic Theory for Principal Component Analysis," *Annals of Mathematical Statistics*, 34, 122–48, 1963.

———, *An Introduction to Multivariate Statistical Analysis*, Wiley Series in Probability and Statistics, 3rd edition, John Wiley and Sons, 2003.

Angeles, G. and Y. You, "Availability of Data for Estimating SES Indices Using Household Surveys," Working Paper, Carolina Population Center, Chapel Hill, NC, 2007.
Babakus, E., C. E. J. Ferguson, and K. G. Jöreskog, "The Sensitivity of Confirmatory Maximum Likelihood Factor Analysis to Violations of Measurement Scale and Distributional Assumptions," *Journal of Marketing Research*, 24, 222–8, 1987.
Bai, J., "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135–71, 1993.
Bartholomew, D. J. and M. Knott, *Latent Variable Models and Factor Analysis, Vol. 7 of Kendall's Library of Statistics*, 2nd editon, Arnold Publishers, London, 1999.
Bollen, K. A., "Multiple Indicators: Internal Consistency or no Necessary Relationship?" *Quality and Quantity*, 18, 377–85, 1984.
———, *Structural Equations with Latent Variables*, Wiley and Sons, New York, 1989.
Bollen, K. A. and K. H. Barb, "Pearson's R and Coarsely Categorized Measures," *American Sociological Review*, 46, 232–9, 1981.
Bollen, K. A. and J. S. Long (eds), *Testing Structural Equation Models*, SAGE Publications, Thousand Oaks, CA, 1993.
Bollen, K. A., J. L. Glanville, and G. Stecklov, "Socioeconomic Status and Class in Studies of Fertility and Health in Developing Countries," *Annual Review of Sociology*, 27, 153–85, 2001.
———, "Economic Status Proxies in Studies of Fertility in Developing Countries: Does the Measure Matter?" *Population Studies*, 56, 81–96, DOI: 10.1080/00324720213796, 2002.
———, "Socioeconomic Status, Permanent Income, and Fertility: A Latent Variable Approach," Working Paper WP-06-90, MEASURE Evaluation Project At Carolina Population Center, Chapel Hill, 2006.
Bouis, H., "The Effect of Income on Demand for Food in Poor Countries: Are Our Food Consumption Databases Giving Us Reliable Estimates?" *Journal of Development Economics*, 44, 199–226, 1994.
Cameron, C. A. and P. K. Trivedi, *Microeconometrics: Methods and Applications*, Cambridge University Press, 2005.
Caudill, S. B., F. C. Zanella, and F. G. Mixon, "Is Economic Freedom One Dimension? A Factor Analysis of Some Common Measures of Economic Freedom," *Journal of Economic Development*, 25, 17–40, 2000.
Choi, I., "Structural Changes and Seemingly Unidentified Structural Equations," *Econometric Theory*, 18, 744–75, 2002.
Conover, W., *Practical Nonparametric Statistics*, Wiley Series in Probability and Statistics, John Wiley and Sons, New York, 1998.
Davis, A. W., "Asymptotic Theory for Principal Component Analysis: Non-normal Case," *Australian Journal of Statistics*, 19, 206–12, 1977.
Deaton, A., *Understanding Consumption*, Oxford University Press, New York, 1992.
Demmel, J. W., *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997.
Deressa, W., A. Ali, and Y. Berhane., "Household and Socioeconomic Factors Associated with Childhood Febrile Illnesses and Treatment Seeking Behavior in an Area of Epidemic Malaria in Rural Ethiopia," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 101, 939–47, 2007.
Diamantopoulos, A. and H. M. Winklhofer, "Index Construction with Formative Indicators: An Alternative to Scale Development," *Journal of Marketing Research*, 38, 269–77, 2001.
Di Bartolo, A., "Human Capital Estimation Through Structural Equation Models with Some Categorical Observed Variables," Working Paper, IRISS At CEPS/INSTEAD, RePEc Handle: RePEc:irs:iriswp:2000-02, 2000.
DiStefano, C., "The Impact of Categorization with Confirmatory Factor Analysis," *Structural Equations Modeling*, 9, 327–46, 2002.
Dolan, C. V., "Factor Analysis with 2, 3, 5 and 7 Response Categories: A Comparison of Categorical Variable Estimators Using Simulated Data," *British Journal of Mathematical and Statistical Psychology*, 47, 309–26, 1994.
Drakos, K., "Common Factor in Eurocurrency Rates: A Dynamic Analysis," *Journal of Economic Integration*, 17, 164–84, 2002.
Fayers, P. M. and D. J. Hand, "Causal Variables, Indicator Variables and Measurement Scales: An Example from Quality of Life," *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 165, 233–61, 2002.
Fernald, L. C. H., "Socio-Economic Status and Body Mass Index in Low-Income Mexican Adults," *Social Science and Medicine*, 64, 2030–42, 2007.
Filmer, D. and L. Pritchett, "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, 1998.

———, "Estimating Wealth Effect Without Expenditure Data—Or Tears: An Application to Educational Enrollments in States of India," *Demography*, 38, 115–32, 2001.
Flury, B., *Common Principal Components and Related Multivariate Methods*, John Wiley and Sons, New York, 1988.
Friedman, M., *A Theory of The Consumption Function*, Princeton University Press, Princeton, NJ, 1957.
Gwatkin, D. R., S. Rustein, K. Johnson, E. A. Suliman, and A. Wagstaff, "Socio-Economic Differences in Health, Nutrition, and Population," Technical Report, World Bank, Volume 1: Armenia–Kyrgyz Republic, 2003a.
———, "Socio-Economic Differences in Health, Nutrition, and Population," Technical Report, World Bank, Volume 2: Madagascar–Zimbabwe, 2003b.
Gwatkin, D. R., S. RutsteIn, K. Johnson, E. Suliman, A. Wagstaff, and A. Amouzou, "Socio-Economic Differences In Health, Nutrition, and Population," Working Papers. Cameroon: WP # 39467; Bangladesh: WP #39465; Burkina Faso: WP # 39466; Colobmia: WP # 39468; Ghana: WP # 39469; Guatemala: WP # 39445; Haiti: WP # 39446; India: WP # 39447; Indonesia: WP # 39448; Kazakhstan: WP # 39449; Kenya: WP # 39470; Malawi: WP # 39450; Mali: WP # 39451; Morocco: WP # 39452; Mozambique: WP # 39453; Namibia: WP # 39454; Nepal: WP # 39455; Nicaragua: WP # 39456; Nigeria: WP # 39457; Peru: WP # 39458; Philippines: WP # 39459; Tanzania: WP # 39471; Turkey: WP # 39460; Uganda: WP # 39461; Zambia: WP # 39463; Zimbabwe: WP # 39464, The World Bank, 2007.
Hand, D. J., *Measurement Theory and Practice*, Kendall's Library of Statistics, Hodder Arnold, 2004.
Harris, D., "Principal Component Analysis of Cointegrated Time Series," *Econometric Theory*, 13, 529–57, 1997.
Hentschel, J. and P. Lanjouw, "Constructing an Indicator of Consumption for the Analysis of Poverty: Principles and Illustrations with Reference to Ecuador," Living Standards Measurement Study Working Paper 124, The World Bank, Washington, DC, 1996.
Hilbe, J. M., "A Review of Stata 9.0," *The American Statistician*, 59, 335–48, 2005.
Hollander, M. and D. A. Wolfe, *Nonparametric Statistical Methods*, 2nd edition, Wiley-Interscience, New York, 1999.
Hong, R. and R. Hong, "Economic Inequality and Undernutrition in Women: Multilevel Analysis of Individual, Household, and Community Levels in Cambodia," *Food and Nutrition Bulletin*, 28, 59–66, 2007.
Horn, R. A. and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1990.
Hotelling, H., "Analysis of a Complex of Statistical Variables into Principal Components," *Journal of Educational Psychology*, 24, 417–41, 498–520, 1933.
Huber, P. J., *Robust Statistics*, John Wiley and Sons, New York, 2003.
Johnson, D. R. and J. C. Creech, "Ordinal Measures in Multiple Indicator Models: A Simulation Study of Categorization Error," *American Sociological Review*, 48, 398–407, 1983.
Johnson, N. L., S. Kotz, and N. Balakrishnan, *Discrete Multivariate Distributions*, JohnWiley and Sons, New York, 1997.
Jolliffe, I. T., *Principal Component Analysis*, 2nd edition, Springer, Heidelberg and New York, 2002.
Jöreskog, K., *Structural Equation Modeling with Ordinal Variables Using LISREL*, Notes On LISREL 8.52, http://www.ssicentral.com/lisrel/ordinal.pdf, 2004.
Judd, K. L., *Numerical Methods in Economics*, MIT Press, Cambridge, MA, 1998.
Kaplan, D., *Structural Equation Modeling: Foundations and Extensions*, Sage Publications, Thousand Oaks, CA, 2000.
Krelle, W., "How to Deal with Unobservable Variables in Economics," Discussion Paper B/414, Bonn University, 1997.
Kullback, S., *Information Theory and Statistics*, Dover Publications, Mineola, NY, 1997.
Lebart, L., A. Morineau, and K. M. Warwick, *Multivariate Descriptive Statistical Analysis*, John Wiley and Sons, New York, 1984.
Maddala, G. S., *Limited-Dependent and Qualitative Variables in Econometrics, Vol. 3 of Econometric Society Monographs*, Cambridge University Press, Cambridge, UK, 1983.
Mardia, K. V., J. T. Kent, and J. M. Bibby, *Multivariate Analysis*, Academic Press, London, 1980.
Maydeu-Olivares, A., "Testing Categorized Bivariate Normality with Two-Stage Polychoric Correlation Estimates," Technical Report, University of Barcelona, Department of Psychology, 2001.
Medina-Solís, C. E., R. P. Núñez, G. Maupomé, and J. F. Casanova-Rosado, "Edentulism Among Mexican Adults Aged 35 Years and Older and Associated Factors," *American Journal of Public Health*, 96, 1578–82, 2006.
Mroz, T. A. and Y. V. Zayats, "Arbitrarily Normalized Coefficients, Information Sets, and False Reports of 'Biases' in Binary Outcome Models," *Review of Economics and Statistics*, 90, 406–13, 2008.

Muirhead, R. J., *Aspects of Multivariate Statistical Theory*, 2nd revised edition, Wiley Series in Probability and Statistics, Wiley-Interscience, New York, 2005.
Muthén, L. K. and B. O. Muthén, *Mplus: Statistical Analysis with Latent Variables, User's Guide*, 3rd edition, Los Angeles, CA, 2004.
Olsson, U., "Maximum Likelihood Estimation of the Polychoric Correlation," *Psychometrika*, 44, 443–60, 1979.
Parlett, B., *The Symmetric Eigenvalue Problem*, Prentice Hall, Englewood Cliffs, NJ, 1980.
Pearson, K., "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, 1901a.
———, "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine*, 2, 559–72, 1901b.
Pearson, K. and E. S. Pearson, "On Polychoric Coefficients of Correlation," *Biometrika*, 14, 127–56, 1922.
Rabe-Hesketh, S., A. Skrondal, and A. Pickles, "Maximum Likelihood Estimation of Limited and Discrete Dependent Variable Models with Nested Random Effects," *Journal of Econometrics*, 128, 301–23, 2005.
Reichlin, L., "Factor Models in Large Cross-Sections of Time Series," Discussion Paper DP3285, CEPR, 2002.
Rencher, A. C., *Methods of Multivariate Analysis*, John Wiley and Sons, New York, 2002.
Scott, C. and B. Amenuvegbe, "Effect of Recall Duration on Reporting Household Expenditures: An Experimental Study in Ghana, Social Dimensions of Adjustment in Africa," Working Paper 6, The World Bank, 1990.
Skrondal, A., "Design and Analysis of Monte Carlo Experiments: Attacking the Conventional Wisdom," *Multivariate Behavioral Research*, 35, 137–67, 2000.
Stata Corp., *Stata Statistical Software: Release 10*, College Station, TX, 2007.
Stock, J. H. and M. W. Watson, "Forecasting Using Principal Components From a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167–79, 2002.
Sumarto, S., D. Suryadarma, and A. Suryahadi, "Predicting Consumption Poverty Using Non-Consumption Indicators: Experiments Using Indonesian Data," *Social Indicators Research*, 81, 543–78, 2007.
Thomas, K. J. A., "Child Mortality and Socioeconomic Status: An Examination of Differentials by Migration Status in South Africa," *International Migration Review*, 41, 40–74, 2007.
Vyas, S. and L. Kumaranayake, "Constructing Socio-economic Status Indices: How to Use Principal Components Analysis," *Health Policy and Planning*, 21, 459–68, 2006.
Webster, T. J., "A Principal Component Analysis of the U.S. News & World Report Tier Rankings of Colleges and Universities," *Economics of Education Review*, 20, 235–44, 2001.
Weisstein, E. W., "Eigenvalue," from MathWorld–A Wolfram Web Resource, http://mathworld.wolfram.com/Eigenvalue.html, 2004.
WIDER, "World Income Inequality Database (WIID)," UNU/WIDER-UNDP World Income Inequality Database, Version 1.0, September 12, 2000, http://www.wider.unu.edu/wiid/wiid.htm, 2000.
Wooldridge, J. M., *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA, 2002.

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix A: The Basic Concepts in Principal Component Analysis
Appendix B: Examples
Appendix C: MIMIC Model
Appendix D: Comparing the Scales in Probit Models

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.