The Use of Discrete Data in PCA: Theory, Simulations, and Applications to Socioeconomic Indices

Stanislav Kolenikov[*] Gustavo Angeles[†]

October 20, 2004

Abstract

The last several years have seen a growth in the number of publications in economics that use principal component analysis (PCA), especially in the area of welfare studies. This paper gives an introduction into the principal component analysis and describes how the discrete data can be incorporated into it. The effects of discreteness of the observed variables on the PCA are overviewed. 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 shed light on some of the issues raised in the theoretical part of the paper. The simulation results show that the currently used method of running PCA on a set of dummy variables as proposed by Filmer & Pritchett (2001) is inferior to other methods for analyzing discrete data, both simple such as using ordinal variables, and more sophisticated such as using the polychoric correlations.

Keywords: welfare indices, principal component analysis, PCA, polychoric correlations, rank correlations, living standards, socio-economic status

JEL classification: C19, C49, I32

[*] Corresponding author. Department of Statistics, CB#3260 University of North Carolina, Chapel Hill, NC 27599-3260, USA, and Centre for Economic and Financial Research, Moscow, Russia. E-mail: skolenik@unc.edu.

[†] Department of Maternal and Child Health and Carolina Population Center, University of North Carolina at Chapel Hill, CB#8120 University Square East, Chapel Hill, NC 27516-3997, USA. E-mail: gangeles@email.unc.edu. The views represented in the paper are those of the authors, and do not reflect the views of Carolina Population Center or US Agency for International Development.

1 Introduction

One of the recurrent ideas and needs in the development economics studies at the micro level is to assess the socio-economic status (SES) of a household or an individual. Such estimates usually serve as an input to another analysis such as inequality or poverty analysis, tabulation of population characteristics by quintiles or deciles, or regressions that involve welfare as an explanatory or dependent variable and aim at explaining the household health status or certain behavior.

Broadly speaking, the socio-economic 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 the socio-economic status in this paper.

Often, straightforward numeric measures of welfare such as household income or consumption are not available or not reliable, especially in non-market economies where a large fraction of economic activities is carried out outside of the market. This does not have to be an illegal black market activity. Family farming where family members are not paid salary, but rather consume a large portion of their produce is a good example.

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, and 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?” Those variables with a small number of clear response categories suffer much less measurement and reporting error than does income or expenditure, although they would still contain a lot of measurement error in terms of measuring socio-economic status.

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 that stand between the proxy and the concept it is supposed to measure. Fortunately, the researcher can observe the possession by a household of several durable goods, as well as record the characteristics of a dwelling. That way, something between 10 and 20 characteristics (possibly with several levels) 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; or from other considerations, such as putting all coefficients to one. That way, the researcher obtains a univariate measure of welfare. It may not have a direct interpretation, say, in dollar terms, unless some measures of income, expenditure or wealth are used explicitly in the analysis as the nominal anchors, and have the coefficients set to one. Otherwise such a measure cannot be used directly for poverty analysis in terms of relating somebody's disposable resources to an absolute figure like $1 a day, but it finds use in ranking individuals, make decisions regarding the allocation of projects that are to benefit the poor, or as an input to other research problems where the researcher is interested in relation between SES and observed behaviors.

A thorough review of the existing methods of SES assessment in application to the fertility studies is given in Bollen, Glanville & Stecklov (2001) and Bollen, Glanville & Stecklov (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". Upon providing a thorough review of the methods and concepts related to SES, such as Friedman's permanent income thesis, they compare the performance in terms of external validity (the explanatory power in a regression with fertility as a dependent variable) of a simple sum of the assets (i.e. total number of durable goods possessed by the household), sum of current values (as assessed by the household itself), sum of median values (where the median value of the asset across all households is taken as the market price of an item), principal components, as well as 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 a simple sum of asset indicators.

The principal component analysis was developed in early 20th century (Pearson 1901*b*, 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 a standard multivariate technique described in such textbooks as Anderson (2003), Mardia, Kent & Bibby (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, Zanella & Mixon 2000), panel data (Bai 1993, Reichlin 2002), forecasting (Stock & Watson 2002), simultaneous equations (Choi 2002) and economics of education (Webster 2001). See also reviews of the factor models in Bai (1993) and Wansbeek

& Meijer (2000). Krelle (1997) gives a review of a number of methods aimed at estimation of unobservable variables, including PCA.

One of the earliest and most influential papers in development economics and population studies for the construction of socio-economic indices that used PCA was Filmer & Pritchett (2001) (and an earlier working paper version Filmer & Pritchett (1998)). They used the data on household assets (primary importance durable goods such as clock, bicycle, radio, television, sewing machine, motorcycle, refrigerator, car), type of access to hygienical facilities (sources of drinking water, types of toilet), number of rooms in dwelling, and construction materials used in the dwelling.

The methodology was quickly accepted by the World Bank (Gwatkin, Rustein, Johnson, Suliman & Wagstaff 2003*a*, Gwatkin, Rustein, Johnson, Suliman & Wagstaff 2003*b*) and ORC/Macro Demographic and Health Surveys[1] as the way to assess socio-economic status of a household based on the household assets (electricity, radio, television, telephone, refrigerator, bicycle, motorcycle, car or truck) and facilities (source of drinking water, toilet type, source of heat for cooking, materials used for flooring, walls, and roofing).

Despite the nice title "Estimating wealth effect without expenditure data — or tears", the paper by Filmer & Pritchett (2001) has not quite solved all of the methodological problems. The primary critique that can be raised for the method used in the aforementioned papers is that the use of dummy variables in the PCA is not justified, as PCA "as is" is only suitable for continuous data. It was developed for the samples from multivariate normal distribution (Hotelling 1933, Anderson 2003, Mardia et al. 1980), and most of the theoretical results, including the implicitly used consistency of the estimates of the factor loadings, were derived under the normality assumption. See Appendix A for technical results.

In fact, the Filmer & Pritchett (2001) go further than just using the discrete welfare indicators as if they were continuous. Instead, a dummy variable was used 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; see argument about numerical stability in Appendix A). The reason for doing so 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

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

treated as *dependent* ones. The analysis must be modified accordingly, since the assets used in the PCA are indicators, or outcomes, of the welfare, rather than determinants of it.

One consequence of using the dummy indicators in PCA for construction of the welfare indices is that doing so introduces a lot of spurious correlations if there are more than two categories for a variable (and thus more than one dummy variable per categorical factor is created). 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 starts "getting confused" as to whether the main source of the common variation in the data is due to the correlation with the unobserved welfare (as we want it to be, and we want PCA to capture this cross-dependence), or due to the correlation among the variables that belong to the common categorical variable (as introduced by the researcher through the use of dummy indicators). Even if the former is a strong relation, it might get blurred by the latter, and thus these spurious correlations may tend to generate incorrect estimates of the socioeconomic index. The goodness of fit measures are going to deteriorate, too, since the PCA will see a noisier covariance matrix.

Besides, 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. We tend to think, however, that if you have additional information, it can and should be incorporated, as it helps producing more accurate results. As it is always the case in statistics and econometrics, the model-based methods produce more efficient results for well-specified models than semi- or non-parametric methods. Here, the weakest form of the model assumptions used is that the researcher can provide ordering of categories based on the substantive knowledge of the problem.

There is a substantial literature on the use of discrete data in mutlivariate methods. The foundations for the use of ordinal data and the foundations of the principal component analysis were developed at the same time and by the same person. The latter was done in Pearson (1901*b*), while Pearson (1901*a*) 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 & Pearson (1922) and Olsson (1979) introduced concepts of *polychoric* and *polyserial* correlations as the maximum likelihood estimates of the underlying correlation between the unobserved normally distributed continuous variables from their discretized versions. Other literature such as Bollen & Barb (1981), Babakus, Ferguson, Jr. & Joereskog (1987), Dolan (1994), and DiStefano (2002), among others, has looked at the effects of categorization in a closely related area of structural equation modelling with latent variables, also known as linear structural relations. We have found only one application of the polychoric correlations in economic indexing systems (Bartolo 2000).

Yet another aspect of PCA that we shall only briefly mention in passing is the effect of complex sample design on the principal component analysis. Skinner, Holmes & Smith (1986) show that analysis that does not take into account the design leads to biased estimates for disproportionate designs (i.e., those where probabilities of selection differ for different members of the population).

The goal of this paper is to provide an overview of PCA and examine how discrete data can be appropriately be used with it. The other purpose of the paper is to examine the performance of different procedures for using PCA with discrete data typically available in household health surveys for measuring SES. We present a number of small analytical examples that demonstrate the results of PCA with simple discrete distributions. For more complex situations, we designed and carried out a large simulation project that also confirmed that the Filmer-Pritchett procedure yields results inferior to other methods.

The remainder of the paper is organized as follows. The next section reviews the main procedures of the principal component analysis, including the general formulation (section 2.1); the specific features of the discrete data that make it more difficult to carry out the PCA with such data (section 2.2); and the definitions and properties of the polychoric and polyserial correlations (section 2.3). Then section 3 introduces our Monte Carlo study. Section 3.1 presents the setup of the simulations comparing the polychoric correlation and related approaches, and the Filmer-Pritchett approach of PCA on dummy variables. Section 3.2 presents the numeric findings, and section 3.3 demonstrates some of them visually. Section 4 concludes.

2 PCA and related procedures

2.1 Principal component analysis

If $\mathbf{x}$ is a random vector of dimension p with finite $p \times p$ variance-covariance matrix $\mathbb{V}[\mathbf{x}] = \Sigma$, then the *principal component analysis* (PCA) solves the problem of finding the directions of the greatest variance of the linear combinations of x's. In other words, it seeks the orthonormal set of coefficient vectors $\mathbf{a}_1, \ldots, \mathbf{a}_k$ such that

$$
\begin{aligned}
\mathbf{a}_1 &= \arg \max_{\mathbf{a}:\|a\|=1} \mathbb{V}\big[\mathbf{a}'\mathbf{x}\big], \\
&\dots \\
\mathbf{a}_k &= \arg \max_{\substack{\mathbf{a}:\|\mathbf{a}\|=1, \\ \mathbf{a}\perp\mathbf{a}_1,\dots,\mathbf{a}_{k-1}}} \mathbb{V}\big[\mathbf{a}'\mathbf{x}\big], \\
&\dots
\end{aligned} \tag{1}
$$

The maxima are those of a convex function on a compact set, and thus exist, and are unique if there are no perfect collinearities in the data, up to the change of the sign of all elements of $\mathbf{a}_k$. The linear combination $\mathbf{a}_k'\mathbf{x}$ is referred to as the k-th *principal component* (PC).

The motivation behind this problem is that the directions of greatest variability give "most information" about the configuration of the data in multidimensional space[2]. The first principal component will have the greatest variance and extract the largest amount of information from the data; the second component will be orthogonal to the first one, and will have the greatest variance in the subspace orthogonal to the first component, and extract the greatest information in that subspace; and so on. Also, the principal components minimize the L_2 norm (sum of squared deviations) of the residuals from the projection onto linear subspaces of dimensions 1, 2, etc. The first PC gives a line such that the projections of the data onto this line have the smallest sum of squared deviations among all possible lines. The first two PC define a plane that minimizes the sum of squared deviations of residuals, and so on.

The principal components analysis can be carried out for both for the theoretical distributions and the actual data. In the latter case, one would analyze the empirical covariance matrix. Plotting several first components against each other can often give good insight into the structure of the data, presence of clusters, nonlinearities, outliers, etc.

There are a number of practical choices that researchers have to make when performing the principal component analysis. The first one is to choose what variables to include in the analysis. The desirable choice is that all variables describe a common

[2] This statement makes exact sense when the vector $\mathbf{x}$ has multivariate normal distribution, and the information is the Kullback-Leibler information (Kullback 1997) between the original distribution, and the joint distribution of the first, …, k-th components.

phenomenon, and the primary application we are looking at in this paper is the welfare analysis. As far as PCA was originally developed for the multivariate normal distribution and samples from it, the PCA will work best on the variables that are continuous and at least approximately normal.

Another important choice to be made is whether the data need to be standardized. If the original variables have wildly different scales, then the PCA will simply pick the variable that has the highest variance as the direction of the greatest variability. Most of the time, the researcher rather wants to find relations between the variables, and for that purpose, one should analyze the standardized data, when each variable has mean zero and variance 1, so that the principal component analysis indeed analyzes the dependencies among the variables rather than differences in measurement scales. The analysis of the standardized data is equivalent to analysis of the correlation matrix of the original data. This would be the default option for most statistical packages, too.[3]

The solution to equation (1) is found by solving the *eigenproblem* for the covariance (or correlation) matrix Σ: find λ's and a's (with an identification condition $\|a\| = 1$) such that

$$\Sigma a = \lambda a \tag{2}$$

Some background on eigenproblems is provided in Appendix B.

The solution of the eigenproblem (2) for the covariance or more commonly correlation matrix gives the set of principal component weights $\mathbf{a}$ (also referred to as *factor loadings*), the linear combinations $\mathbf{a}'\mathbf{x}$ (referred to as *scores*) and the eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p$. It is easy to establish that $\mathbb{V}\big[\mathbf{a}'_k\mathbf{x}\big] = \lambda_k$ given that $\mathbb{V}\, x_j = 1$ (which would be the case of the standardized data, or correlation matrix), so that the eigenvalues are the variances of the corresponding linear combinations. Then the linear combination that corresponds to the largest eigenvalue is the one that has the greatest variance. Note that the principal components are only defined up to a sign, so in the applied work it is always worth checking if the principal components correspond to the desired direction of the feature variation, such as higher values of the score should represent richer households. Quite often the first component can be interpreted as a measure of size, or a degree of expression of a certain feature, while the second, the third, and so on components might have an interpretation of some structure of that feature.

[3] See also discussion in Appendix E. Anderson (1963) finds deriving asymptotic results for the PCA on a correlation matrix more difficult, and attributes that to the "difficulties of interpretations for correlations" as compared to the interpretation of covariances.

A popular measure of fit by the principal components is referred to as *proportion of explained variance*:

$$R_k^{\text{ind}} = \frac{\lambda_k}{\lambda_1 + \ldots + \lambda_p} \tag{3}$$

or

$$R_k^{\text{cum}} = \frac{\lambda_1 + \ldots + \lambda_k}{\lambda_1 + \ldots + \lambda_p} \tag{4}$$

Most of the time, one wants to see high proportions of variance being explained by the first principal component, or few first components. It means that the variables used as an input have a lot in common. Also, those proportions of explained variance find a use in determining the number of “significant” components. For more details, see Appendix A.

The PCA is intrinsically a linear procedure, so it is non-robust, in the sense of Huber (2003), to various distributional assumptions violations. In particular, if the distribution of x exhibits high skewness and/or kurtosis, the weights and eigenvalues in PCA will have higher variances, and converge to their asymptotic distributions slower (Davis 1977).

The model implicitly assumed by treating the first principal component as a suitable index of welfare can be stated as follows. Suppose there is an unobserved variable ξ with a number of observed indicators (such as quality of water or housing materials; we shall postpone the treatment of complications arising because of discreteness until section 2.2):

$$x_k = \Lambda_k \xi + \delta_k, \quad k = 1, \ldots, p \tag{5}$$

This is the simplest model of the *confirmatory factor analysis* (CFA), which in turn is a part of the *structural equations modelling* framework (Bollen 1989, Bollen & Long 1993, Bartholomew & Knott 1999, Kaplan 2000). More general CFA models may include several factors ξ, possibly correlated; allow different factors to be loaded onto different variables; or allow the measurement errors δ be correlated across different indicators x. Note again that in this context, the (unobserved) level of welfare ξ is the explanatory variable, and the observed x’s are the dependent variables in the system of regression equations (5).

2.2 Discrete data

A practically important violation of the normality assumption underlying the PCA occurs when the data are discrete. There are several kinds of discrete data one can encounter in empirical analysis. 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 such that there is a monotone relation between them. The examples might be: different levels of education (no education, primary, secondary, higher, professional or advanced degree), subjective well-being on a ladder scale (from 1 to 9 where 1 is the most miserable person, and 9 is the happiest one), or, in the context of the welfare studies we were motivated by, different construction materials used in the building (no roof is worse than a straw roof, which in turn is worse than than a wooden roof, and all of those are dominated by a cement 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 *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.

In what follows, we shall only concentrate on the ordinal data (including binary data). For the analysis of the truly categorical data with no natural ordering of categories, there is no obvious advice as to what should be done. We argue in this paper that the Filmer-Pritchett PCA procedure of generating dummy variables from such data introduces undesirable correlations. Other solutions can be sought in the framework of the structural equation models, as explained in the end of the previous section. If the categorical factors such as region are believed to be the *determinants* rather than *outcomes* of the welfare, then the applicable model is referred to as a *MIMIC* (multiple indicators and multiple causes) model. Discussion of this model is beyond the scope of this paper. An interested reader may be referred to the standard literature on structural equations (Bollen 1989, Bollen & Long 1993, Bartholomew & Knott 1999, Kaplan 2000).

When we turn to the ordinal data, a simple and naïve 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[4]. If the ordinal data are used as if they were

[4] In the PCA case however one can find an additional justification for this approach by noting that using ordered categories can be viewed as computing Spearman's rank correlation ρ_S instead of Pearson's moment correlation in the analysis. 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, Morineau & Warwick 1984, Sec. 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 unless transformed. They are also robust to outliers which may not be much of an issue for discrete variables, but may be an issue for skewed distributions such as that of raw income data. See Appendix C for some details on rank correlations.

continuous, problems may arise. The violations of the distributional assumptions in PCA incurred by the 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 (5), the observed indicators really are the dependent variables!

The most common way of analyzing the discretization effects is to assume that there are underlying continuous variables x_k^* that have the pre-specified relation to the underlying latent variable (welfare) ξ, as in the CFA model (5):

$$\mathbf{x}^* = \Lambda_x \xi + \delta, \quad \mathbb{V}[\delta] = \Theta_\delta = \operatorname{diag}[\delta_1, \ldots, \delta_K], \quad \mathbb{V}[\xi] = \Phi \tag{6}$$

If the observed x_k's are ordinal with the categories $1, \ldots, K_k$, then it is assumed that they are obtained by discretizing the underlying x_k^* according to the set of thresholds $\alpha_{k1}, \ldots, \alpha_{k,K_k-1}$:

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

where $\alpha_{k,0} = -\infty$, $\alpha_{k,K_k} = +\infty$. So, x's are dependent variables, and if ξ's were observed, we would use ordered dependent variable models to analyze the relations between ξ and x.

Let us illustrate the above notation with a simple example. It will also show some of the problems that arise because of the discrete character of the data, such as excessive skewness and kurtosis.

Example 1.

Table 1: The joint distribution of the discretized data in Example 1.

	$x_1 = 1$	$x_1 = 2$	$x_1 = 3$	$x_1 = 4$	Marginal
$x_2 = 3$	0.2%	2.1%	6.9%	6.7%	15.9%
$x_2 = 2$	0.8%	8.2%	20.6%	14.4%	44.0%
$x_2 = 1$	1.3%	10.1%	19.0%	9.7%	40.1%
Marginal	2.3%	20.4%	46.5%	30.8%	100%

On Fig. 1, the parameters are as follows: $\alpha_{1,1} = -2$, $\alpha_{1,2} = -0.75$, $\alpha_{1,3} = 0.5$; $\alpha_{2,1} = -0.25$, $\alpha_{2,2} = 1$, and the correlation of the underlying bivariate normal is 0.2. The cell proportions and the marginal are given in Table 1. Note that the joint distribution of x_1 and x_2 is an example of the opposite skewness: x_1 is skewed to the left (with skewness -0.40), while x_2 is skewed to the right (with skewness 0.39). Another important feature of this discretized bivariate data is high kurtosis: the kurtosis of x_1 is 2.52, and kurtosis of x_2 is 2.04. Typically,

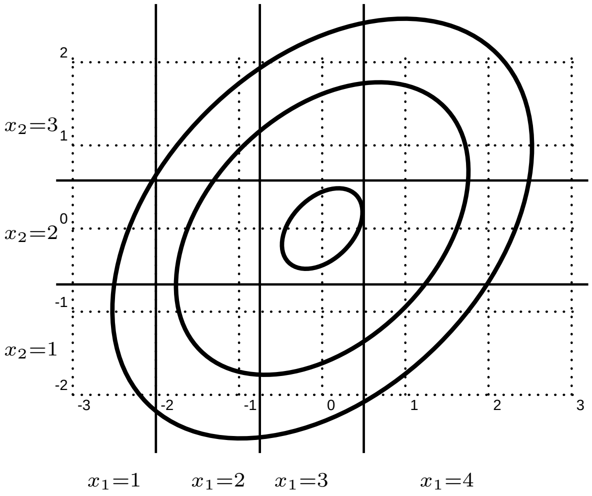


Figure 1: Categorized bivariate normal distribution.

high kurtosis is associated with the data that have heavy tails at infinity; this example, however, shows that even despite the finite range, the kurtosis is substantial. ⊠

There is a number of implications of the discrete character of the data if the observed discrete x_k's are used directly in the standard principal component analysis. The problems related to the discrete data have received a considerable attention in quantitative sociology (Olsson 1979, Bollen & Barb 1981, Johnson & Creech 1983, Babakus et al. 1987, Dolan 1994, DiStefano 2002).

First, the distributional assumptions (normality) are seriously violated. Obviously, the discrete data do not have a density (at least with respect to Lebesgue measure). 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. Even with moderate discretization in Example 1, the skewness and kurtosis were not negligible. PCA only addresses the second moments of the data, in essence approximating the real data with the normal distribution of the same mean and covariance matrix.

As far as the distribution of the data is not a multivariate normal, the standard results on the asymptotic distributions of the eigenvalues and eigenvectors[5] need to be modified (Davis 1977). The normality will still hold as the eigenvalues and eigenvectors

[5] See Appendix A for the existing results on the asymptotic distributions of the eigenvalues and eigenvectors.

are functions of the covariance matrix, whose entries are asymptotically normal since they are sums of i.i.d. variables. The parameters of the resulting asymptotic normal distribution, however, depend crucially on the fourth moments of the data generating distributions.

Second, and maybe even more important, consequence of the discreteness is that the covariances or correlations between the discretized versions x_1, x_2 of the variables of interest are not equal to the "true" covariances or correlations of the (unobserved) underlying variables x_1^*, x_2^*. They mostly tend to be biased towards 0, as the following example demonstrates.

Example 2. Suppose two binary variables x_1, x_2 were obtained by categorizing x_1^*, x_2^* that came from a bivariate normal distribution with standard normal marginals. Denote $\mathrm{corr}(x_1^*, x_2^*) = \rho$, and

$$\Phi_2(s,t;\rho) = \int\limits_{-\infty}^{s}\int\limits_{-\infty}^{t} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\Bigl[-\frac{1}{2(1-\rho^2)}\bigl(u^2 - 2\rho uv + v^2\bigr)\Bigr]\,du\,dv \qquad (8)$$

the cdf of the bivariate standard normal distribution. If the thresholds are given by $\alpha_{1,1}$ and $\alpha_{2,1}$ ($\alpha_{i,0} = -\infty$, $\alpha_{i,2} = +\infty$, $i = 1$, 2), then the proportions in cell (i,j) is

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

where $i,j = 1,2$. Then it is easy to establish that

$$\begin{gathered}\bar{x}_1 = \pi_{10} + \pi_{11}, \quad \bar{x}_2 = \pi_{01} + \pi_{11} \\ \mathbb{V}\,x_1 = (\pi_{10} + \pi_{11})(\pi_{01} + \pi_{00}), \quad \mathbb{V}\,x_2 = (\pi_{01} + \pi_{11})(\pi_{10} + \pi_{00}), \\ \mathrm{Cov}[x_1, x_2] = \pi_{00}(\pi_{10} + \pi_{11})(\pi_{01} + \pi_{11}) - \pi_{01}(\pi_{10} + \pi_{11})(\pi_{00} + \pi_{10}) - \\ -\pi_{10}(\pi_{00} + \pi_{01})(\pi_{01} + \pi_{11}) + \pi_{11}(\pi_{00} + \pi_{01})(\pi_{00} + \pi_{10})\end{gathered} \qquad (10)$$

The dependence of the observed correlation on the underlying correlation and the threshold structure in a 2×2 case is shown on Fig. 2. The lines from top to bottom are based on thresholds of $\alpha_{1,1} = 0$ and $\alpha_{2,1} = 0$ (and hence marginal proportions of 0's and 1's equal to a half, labelled as "Half-half" on the plot), 0.67 and 0.67 (which gives the proportion of 0's about 3/4, and proportion of 1's about 1/4, labelled as "Upper Q - Upper Q" on the plot), 0 and 0.67 (labelled as "Half - Upper Q"), and -0.67 and 0.67 which gives the opposite skewness (labelled "Upper Q - Lower Q"). All of the lines lie below the diagonal which is in agreement with the above suggestion that the correlations are biased towards zero. If the thresholds are not the same, the observed correlations do not reach 1 even if the underlying correlation is 1. The worst is the case of opposite skewness of the binary indicators: the correlations then do not exceed 0.33. ☒

Example 2 shows that in the extreme case of dichotomizing the continuous distribution, the correlations are largely underestimated. Even if the underlying variables x_1^*, x_2^* are perfectly related (i.e., the correlation between them is 1), their discretized manifestation may show correlation that is far from 1 unless the categorization thresholds match exactly. For all values of the correlation, however, the correlation based on the discrete versions x_1, x_2 is lower than the correlation of the underlying "starred" variables. In a more general case of more than two categories, categorization can be viewed as a measurement error with nonlinear properties, and the authors are not aware of the general literature that shows that correlations go down because of discretization.

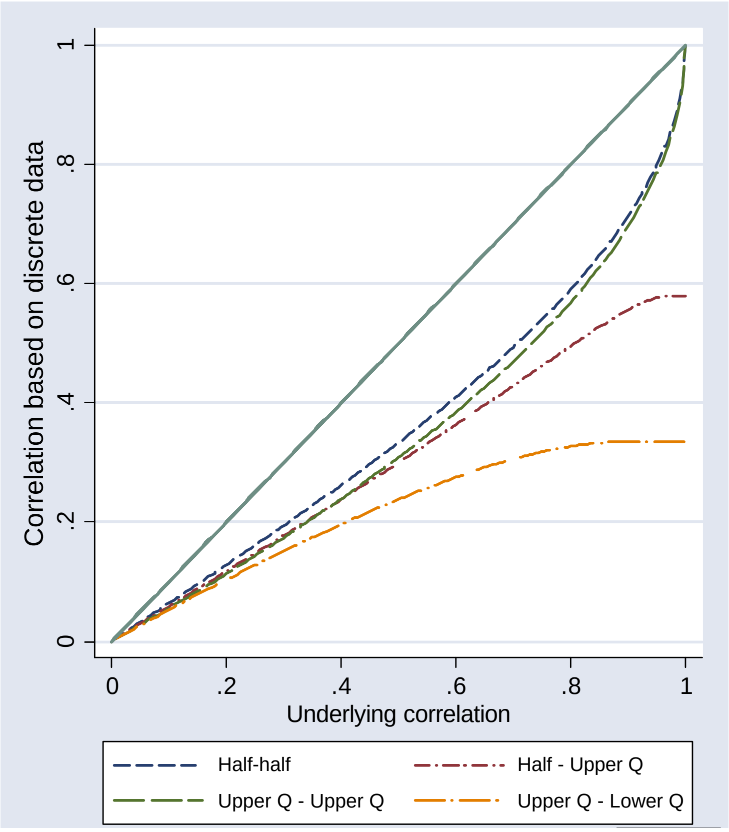


Figure 2: Correlation of the bivariate binary data obtained from a bivariate normal distribution.

It is however very well established empirically in sociological literature cited above.

As long as the covariance structure of the observed variables no longer complies with the theoretical model like (5), the estimated principal component weights will also be biased and inconsistent. Thus the quality of approximation of the first principal component as an index of welfare in economic applications might be dubious.

Also, the PCA run on the discretized x's underestimates the share of variance that falls onto the first few components (see Example 3 in Appendix B). The distortions are greater with the fewer number of categories per variable (5 or more categories are considered sufficient for the distortions to be negligible), higher skewness and kurtosis of the discrete variable distributions (which is the case when most of the data are concentrated in a single category), and opposite skewness of different categorical variables, as in Example 1).

The use of the dummy variables corresponding to individual categories of x_k leads to further violations of the model assumptions, as was argued in Section 1. It introduces spurious correlations effectively smearing the dependence on a few factors across the whole covariance matrix, and at the same time reducing further the proportion of variance explained by the first few components. Also, all the information about the natural ordering of the categories is lost, and that may lead to a substantial distortion of the results.

Two analytical examples are worked out in Appendix D. Example 4 shows some of the possible consequences of discretization. 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, so all of the variation can be explained with a single score. If additionally some categories are about equally populated, then the principal component may not be well defined, which would result in a high sampling variation of it. In the extreme case of all categories having the same proportions, any weights that sum up to zero may serve as a valid first principal component. The empirical implication of this is that the first PC will be highly unstable and wiggle due to sampling fluctuations that would make some categories more populated.

The algebra is further developed through in Example 5 that 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 category with the largest number of observations, and the second largest weight of a different sign, to the second largest category. The direction of the greatest variability is the one that "connects" the two largest clusters. Indeed, those are the variables that have the largest variability, and define the largest off-diagonal entry of the correlation matrix. Thus the combination of weights that gives larger weight to those categories (and are of different signs since the correlation in question is negative) will produce larger variance. This also seems to be 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.

Finally, the natural ordering of categories is not generally reproduced by the principal component analysis, so the only condition that identifies the ordering would 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, although even the binary ownership indicators tend to produce reasonable results in practice. Otherwise, unless the two largest categories are the poorest and the richest members of the population, the first principal component would fail to give a meaningful direction of the welfare change, and the scores with low counts will not be well reproduced.

2.3 Polychoric and polyserial correlations

This section introduces an alternative approach to the analysis of the discrete data in PCA, in particular, to computing the correlations between two ordinal variables. The approach originated in Pearson (1901*a*) and was further developed in Pearson & Pearson (1922) and Olsson (1979). In fact, it is very similar to the assumptions one makes in deriving an ordered probit model (Maddala 1983, Wooldridge 2002). A general treatment is given in Jöreskog (2004*b*) for LISREL software (SSI 2004) that he originated, and that remains the leader in the multivariate analysis of the ordinal variables.

Suppose two ordinal variables x_1, x_2 are obtained by categorizing two variables x_1^*, x_2^* with distribution

$$\begin{pmatrix} x_1^* \\ x_2^* \end{pmatrix} \sim N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \quad -1 \le \rho \le 1 \tag{11}$$

The categorizing thresholds for the two variables are given by $\alpha_{1,0} = -\infty < \alpha_{1,1} < \ldots < \alpha_{1,K_1} < \alpha_{1,K} = \infty$, $\alpha_{2,0} = -\infty < \alpha_{2,1} < \ldots < \alpha_{2,K_1} < \alpha_{2,K} = \infty$, so that $x_i = k$ when $\alpha_{i,k-1} < x_i^* \le \alpha_{i,k}$, $i = 1, 2$. Then the theoretical proportions of the data in each cell can be found as (9). (See also Example 1.)

Assuming that observations are i.i.d., the likelihood can be written down as

$$L(\rho) = \prod_{i=1}^{N} \prod_{m=1}^{K_1} \prod_{l=1}^{K_2} \pi_{(}m, l; \rho, \alpha)^{I(x_{i,1}=m, x_{2,i}=l)} = \prod_{i=1}^{N} \pi(x_{i,1}, x_{i,2}; \rho, \alpha) \qquad (12)$$

$$\ln L = \sum_{i=1}^{N} \ln \pi(x_{i,1}, x_{i,2}; \rho, \alpha) \qquad (13)$$

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

$$\alpha_{i,j} = \Phi^{-1}\Bigl(\frac{-1/2 + \#\{x_i \leq j\}}{N}\Bigr), \quad j = 1, \dots, K_i, \qquad (14)$$

Second, the correlation coefficient is estimated by maximizing (13) 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 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 impractical.

In the multivariate setting with more than two variables, the estimate of the correlation matrix is obtained by combining the pairwise estimates of the polychoric correlation in the third stage of the estimation procedure. Jöreskog (2004*a*) refers to methods of this type as "bivariate information maximum likelihood" (BIML). The full information likelihood estimates obtained by writing out the full multivariate likelihood and maximizing it over the thresholds and correlation coefficients may be more advantageous theoretically, but hardly feasible technically. It is not guaranteed that the resulting estimated correlation matrix is non-negative definite, however, in neither of the two cases.

The assumption of normality can be tested by looking at the proportion of the data in each cell and comparing it to those under normality, with estimated thresholds and the polychoric correlation coefficient. The first test is the likelihood ratio test of the saturated model that does not make any distributional assumptions (cell proportions) and the normality-implied one:

$$LR = -2\Big(\sum_{m=1}^{K_1}\sum_{l=1}^{K_2} n_{ml} \ln \frac{n\pi(m,l;\hat{\rho},\hat{\alpha})}{n_{ml}}\Big) \tag{15}$$

where $n_{ml} = |\{i : x_{i,1} = m, x_{i,2} = l\}|$ is the number of observations identified by m, l-th categories of variables x_1 and x_2. The second test is Pearson goodness of fit test for distributions:

$$X^2 = \sum_{m=1}^{K_1}\sum_{l=1}^{K_2} n_m l \frac{(n_{ml}/n - \pi(m,l;\hat{\rho},\hat{\alpha}))^2}{\pi(m,l;\hat{\rho},\hat{\alpha})} \tag{16}$$

Both of those statistics would have an asymptotic χ^2 distribution with $K_1K_2 - K_1 - K_2$ degrees of freedom.

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 x_1 with underlying standard normal x_1^* discretized according to the thresholds $\alpha_{1,0} = \infty < \alpha_{1,1} < \ldots < \alpha_{1,K_1} < \alpha_{1,K} = \infty$, and the continuous variable x_2 (assumed to have the standard normal distribution) is as follows:

$$\begin{aligned} L(\rho,\alpha; x_1 = k, x_2) = f(x_1 = k, x_2; \rho, \alpha) = \mathrm{Prob}[\alpha_{1,k-1} < x_1^* \le \alpha_{1,k}|x_2]\phi(x_2) = \\ = \big(\Phi(\alpha_{1,k} - \rho x_2) - \Phi(\alpha_{1,k-1} - \rho x_2)\big)\phi(x_2) \end{aligned} \tag{17}$$

as long as $\mathbb{E}[x_1^*|x_2] = \rho x_2$. 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 obtain the polyserial correlation between the two variables.

Having estimated the correlations, one can proceed to the PCA in the standard manner, i.e., by solving the eigenproblem for the estimated correlation matrix.

One of the authors has developed a module for Stata software (Corporation 2003) for the polychoric correlation analysis. It can be found from within Stata connected to the Internet by typing

```
findit polychoric
```

in Stata command prompt.

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

$$\mathbb{E}[x^*|x = j] = \int_{\alpha_{j-1}}^{\alpha_j} u\phi(u)\,du = \phi(\alpha_{j-1}) - \phi(\alpha_j), \quad \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad (18)$$

This value can be used instead of $x = j$ 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 is the same as the distance between the categories 2 and 3, or 3 and 4.

3 Monte Carlo study

3.1 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 development economics as in Filmer & Pritchett (2001), i.e., for ranking households by their welfare. The main theme of the simulation was to set up a model of the form (5) 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.

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% (1 discrete, 1 continuous) to 100%.
- The distribution of the underlying factor: normal; uniform; lognormal; bimodal (a mixture of two normals).
- The proportion of the variance explained: 80%, 60%; 50% if the total number of indicators was greater than 4; 40% and 30% if the total number of indicators was greater than 7.

- The values of Λ: all ones; one or two of the discrete variables have $\Lambda = 3$; one or two of the continuous variables have $\Lambda = 3$; one discrete and one continuous variables have $\Lambda = 3$. (See discussion in Appendix A on implications for PCA.)

- 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 $\mathrm{Prob}[x^* < z] = F(z)$, $u_1, \ldots, u_{K-1} \sim U[0, 1]$, and $u_{(1)}, \ldots, u_{(K-1)}$ is the set of order statistics from $u_1, \ldots, u_{K-1}$, then $\alpha_k = F^{-1}(u_{(k)})$).

- The sample sizes: 100, 500, 2000, 10000.

- 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 & Pritchett (2001); PCA on the ordinal variables with the number of the category replaced by the group means given by (18); 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 selection of a particular combination of the simulation parameters was

$$\mathrm{Prob}[\text{select}|\text{simulation settings}] = \exp(-(3 + 0.25p_d + 0.03p_c)) \qquad (19)$$

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 combinations arising for each extra discrete variable. This sampling procedure resulted in approximately 1% sample of all settings combinations, with the total sample size of 947434 observations, and the sum of weights (the estimate of the total population size) of 99.744 millions. (This would be the total sample size should we run the simulation for each combination of parameters.) Those observations came from 189756 unique samples (combinations of settings). Some observations were lost due to the difficulties with the numeric likelihood maximization 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 (Stata file size is 277 Mbytes).

The simulation was performed on the server of statistical applications at UNC[6] as well as on several personal computers that the authors had access to. The software platform is Stata Special Edition, version 8.2 (Kolenikov 2001, Corporation 2003). The project was spread into 41 separate threads. On average, a thread took about 2 to 4 days on a Pentium IV 1 GHz 256Mb RAM PC (single task), or 5 to 10 days on the multitask server, the workload due to nonlinear maximization involving numerical integration of the bivariate normal probabilities in the maximum likelihood estimation step. Also, some matrix manipulations such as product matrix accumulation might have been pretty long for large sample sizes and large number of variables, especially when the ordinal variables were expanded into the sets of dummy variables. Each run required no more than 10 Mbytes of RAM.

3.2 Results

This section describes the basic analysis of the simulation results. We performed regression analysis with several performance measures and the simulation settings to characterize numerically the differences in the PCA methods for discrete data.

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. As discussed in Section 2.1 and Appendix A, the most popular measure is the proportion of the explained variance.

The external measures of performance are those relating the estimated first PC with “the truth”, i.e. ξ, in the context of applications of SES where the scores are used to classify individuals into quintiles, or other rank groups, used for poverty, service use analaysis and ultimately, for policy advice. We will examine the correlation of rankings produced by different PCA procedures with the underlying score ξ, and compare the quintiles groups produced by the two scores. Thus the first of our measures is the Spearman rank correlation[7] of the empirical first PC with the original factor ξ_1. As discussed in Appendix C, Kendall’s τ might have been a more interpretable measure

[6] See http://www.unc.edu/atn/statistical/. The domain within a Sun E15K server has 20 processors with the clock speed of 1.05GHz, and 40 GB of memory.

[7] The definitions and useful facts regarding the rank correlations are given in Appendix C. Rank correlations show how similar are the rankings of individuals produced by two variables.

Table 2: 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
Theoretical explained proportion		-623.493	1927.794	-1402.261	-1268.915
		(1.469)**	(1.706)**	(1.446)**	(2.450)**
Analysis type					
Original: base	189756	0	0	0	0
		(.)	(.)	(.)	(.)
Filmer-Pritchett	189511	-556.864	-332.814	236.595	243.223
		(2.506)**	(1.963)**	(1.249)**	(2.742)**
Group means	189328	-339.320	-226.089	172.133	143.568
		(1.157)**	(0.976)**	(0.860)**	(1.441)**
Ordinal	189511	-345.287	-231.128	175.534	147.047
		(1.159)**	(0.976)**	(0.863)**	(1.444)**
Polychoric	189328	-157.689	-223.402	169.980	142.112
		(1.125)**	(0.983)**	(0.861)**	(1.440)**
Distribution of ξ					
Normal: base	233688	0	0	0	0
		(.)	(.)	(.)	(.)
Lognormal	237222	-203.238	-734.501	421.746	836.829
		(0.694)**	(0.764)**	(0.622)**	(1.021)**
Bimodal	234612	5.190	-97.136	36.580	209.515
		(0.480)**	(0.690)**	(0.562)**	(1.063)**
Uniform	241912	23.847	65.141	-50.568	47.110
		(0.481)**	(0.642)**	(0.586)**	(1.079)**
Average no. categories					
Filmer-Pritchett	189511	-134.764	0.741	-3.673	7.846
		(0.444)**	(0.375)*	(0.230)**	(0.507)**
Other discrete	568167	15.007	20.040	-15.211	-11.223
		(0.183)**	(0.169)**	(0.148)**	(0.252)**
Log samplesize		-1.420	9.797	-6.202	-10.365
		(0.137)**	(0.171)**	(0.144)**	(0.249)**

Cluster corrected standard errors in parentheses. Other controls include: the threshold structure; the factor loadings; the number of discrete and continuous variables. Total number of observations is 947434. Number of clusters is 189756.

of relation between the two variables, but it is prohibitively computationally intensive to be used in simulations. Two other measures are based on the quintile groups of the theoretical and observed welfare scores. Those are an overall quintile misclassification rate and the misclassification in the first quintile, i.e., the share of observations that originally belonged to the first quintile, but were classified elsewhere by the empirical welfare measure. Other correlations and misclassifications were available in the original data set, too, but not used in the current analysis.

How do we gauge the performance of different PCA procedures? If the welfare measure is to be accurate, then 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.

Table 2 presents the regression results based on the simulations. The operational measures used in the regressions are 1000 times the inverse probit transformation of the previously discussed goodness of fit variables[8]. The factor of 1000 was used to scale up the standard errors in the regressions to become close to 1, thus keeping Table 2 compact and informative. The inverse probit transformation was used to bring the original scale of [0,1] of all four measures to $(-\infty, +\infty)$. The rationale of the transformation is the same as the motivation of the probit model compared to the linear probability model: to combat heteroskedasticity and skewness of the residuals at the extremes of the probability ranges. The inverse probit transformation of the original data allowed also for some standard regression diagnostic tools such as residual plots (not shown here). They mostly showed appropriateness of the model specification, although with somewhat skewed and heavy-tailed distributions of residuals. The excessive skewness and kurtosis was traced to be due to either the lognormal distribution of ξ, or to Filmer-Pritchett procedure (see also the discussion of the coefficients and their significance below). The same regression analysis was repeated with the original data without any transformations, and it revealed strong nonlinearity problems, as well as much higher skewness of the residuals, something to be expected from linear probability models.

[8] So, the second column is $1000 \cdot \Phi^{-1}$(explained variance), the third one, $1000 \cdot \Phi^{-1}$(Spearman correlation), the fourth one, $1000 \cdot \Phi^{-1}$($\#$obs. classified into a quintile different from the original quintile$/N$), and the last one is $1000 \cdot \Phi^{-1}$($\#$obs. from the first quintile of ξ classified elsewhere by the empirical measure$/0.2N$). Informally speaking, in the first two regressions, more is better; and in the last two regressions, less is better.

The list of explanatory variables include the simulation settings and their functions and combinations. The ones not shown in the table are[9]:

- the threshold structure proportions (proportion of discrete variables with bottom, center, top dominated categories, and random thresholds; uniform thresholds is the base)
- the threshold structure dummy variables[10]
- the factor loadings (whether there were any variables with $\Lambda_k = 3$ as opposed to the default $\Lambda_k = 1 \, \forall k$, and whether those variables were discrete, continuous, or both)
- dummy indicators of each particular combinations of discrete and continuous variables.

The first column shows additionally 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 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.

The reported results are for the regressions on all observations in the data set, with probability weights given by the selection probabilities (19), and corrections of the covariance matrix of the estimates by clustering. The latter is necessary as long as the observations based on the same Monte Carlo sample (but different in the type of the analysis performed) are strongly related to each other.

Let us list the findings we consider interesting. A short note on the interpretation of the coefficients may be in place here: the value of the coefficient of 100 means that a unit change in the explanatory variable shifts the linear prediction $x\beta$ by 0.1, and the

[9] The complete tables are available at
`http://www.unc.edu/~skolenik/cpc/polychnoric-technical-regtable.pdf`.

[10] The categories were defined as follows: "mostly uniform": uniform distribution of the thresholds, as explained in Section 3.1, for at least $3/4$ of the discrete variables; "uniform or random": all ordinal variables have either uniform or random threshold settings; "center dominated": in at least one of the variables, the middle category has 50% of population, and the remaining half of observations is distributed uniformly across other categories; it represents high kurtosis with low skewness; "skewed": in at least one of the variables, either the top or the bottom category consists of 50% of population, and the remaining half of observations is distributed uniformly across other categories; it represents both high kurtosis and high skewness; "opposite extremes": two variables have domination in different parts of their distributions, like top- and bottom-dominated, or center- and bottom-dominated; it is known to be the most difficult case for the polychoric coefficient estimation.

ultimate response in its original $[0,1]$ scale by 0.04 near the middle of that range; 0.03 near the points 0.25 and 0.75 (i.e., misclassification rates of 25%, or rank correlation of 0.75, which are rather reasonable values for some of the combinations in our data set); 0.02 near the points 0.1 and 0.9; and 0.1 near the points 0.05 and 0.95.

First, it was very reassuring that relatively few variables (64 in our regressions) yield R^2 above 0.90 (0.81 for the first quintile misclassification rate). Most of the performance of the PCA is thus explained by the factors used in the regression.

Second, the comparison of the different methods is directly accessible through the *Analysis type* block of the table. In all four regressions the Filmer and Pritchett procedure is performing worse than any of the other methods, as evidenced by the largest coefficient across all methods. The baseline for the analysis type was the PCA based on the original unobserved x^*, so the regression coefficients show the deterioration of performance relative to that case. The three other methods based on discrete indicators[11] performed about the same, and the top performing method (but infeasible in the field applications) is the PCA based on the original continuous variables.

Third, we can identify the most important explanatory variables, as evidenced by their t-statistics (not reported in the table; the analysis is based on Stata output). The t-statistics may seem to be too big, for any reasonable standards, but the data came from a controlled experiment, and the data set size is about 1 million observations, so one should not be surprised seeing both tiny standard errors and strong effects. The t-statistics still do convey important information on the relative importance of the variables in the regression.

The most important explanatory variable is the theoretical share of explained variance (the first line of regressor coefficients). This is not surprising, as long as this is the primary variable that controls the closeness of the indicators x^* and x to the underlying welfare ξ. The t-statistics vary between 424 (reported explained variance) to 1130 (rank correlation).

The next important factor is the distribution of the underlying ξ, or rather the fact that it is lognormal (in the *Distribution of* ξ block). The lognormality leads to a substantial deterioration of the performance of the empirical PC. The t-statistics range between 292 (the reported explained variance) and 961 (the rank correlation). We believe this has more to do with the high kurtosis rather than high skewness of this distribution, as another asymmetric distribution used in our analysis, the mixture of two normal distributions, did not produce so bad results. This is also in agreement with the theoretical

[11] I.e., the ordinal PCA based on the ordinal variables scaled to the "standard" Likert scale (1, 2, 3, . . .); the polychoric PCA; and the group means PCA based on the ordinal variables assigned scores based on the underlying standard normal distribution.

results on the PCA in non-normal case (Davis 1977, as also reported in Appendix B).

Those two variables were the two most significant in all four regressions. The next group of variables that came up among the most significant ones are the analysis type indicators (see the *Analysis type* block of coefficient estimates). The coefficients in the table are the differences in performance from the PCA based on the original (unobserved) continuous data. For the ordinal or group means analysis, the t-statistics were around 300 in the explained variance regression, around 230 in the rank correlation regression, around 200 in the overall misclassification rate regression, around 100 in the first quintile misclassification rate regression. The t-statistics of the polychoric PCA analysis were lower at 140, 227, 197, and 99, respectively, so the real difference is only in the share of explained variance that is estimated consistently by polychoric PCA, but not other methods. The polychoric PCA tends to produce the results closer to the benchmark PCA on the original variables, while the standard errors are essentially the same as for the ordinal or group means analysis. For the Filmer-Pritchett procedure, the t-statistics were 222 in the explained variance regression, 190 in the overall misclassification rate regression, 170 in the rank correlation regression, and 90 in the Q1 misclassification rate regression. Even though the coefficients of the Filmer-Pritchett procedure dummy variable is greater in absolute value than those for other discrete data procedures, they also have standard errors that are about twice as large as other discrete methods. This is an indication that not only the Filmer-Pritchett procedure gives worse results on average, but also that it is less stable in performance, with the residual variance being notably greater for the Filmer-Pritchett subsample.

The interaction of the number of categories and the Filmer-Pritchett procedure indicator (in the block *Average number of categories*, bottom of the first page of the table) also came out to be the second most significant variable in the explained variance regression ($t = 303$, more categories leads to smaller explained variance). A possible interpretation of this can be proposed as follows: the number of variables in the denominator of (3) in the Filmer-Pritchett procedure increases, while the information they can explain, or variability of the first PC they can produce, remains the same.

The lack of information about ξ contained in a small number of variables is also an important factor for the external measures. The first line at the third page of the table shows that the extreme case of 2 discrete variables and 0 continuous variables yields t-statistics of 274 for the rank correlation regression, 255 for the overall misclassification rate regression, and 83 in the Q1 misclassification rate regression.

Finally, the bimodal distribution dummy was one of the most significant regressors in the Q1 misclassification rate regression, with $t = 197$.

The comparison of the absolute values of the coefficients in order to assess the mag-

Table 3: Monte Carlo simulation results: Number and type of variables.

	Share of explained variance	Rank correlation with ξ	Overall misclassification rate	Misclassification in Q1
Largest difference	$\langle 1,1\rangle$ vs. $\langle 11,0\rangle$ 318.93 (4.18)	$\langle 2,0\rangle$ vs. $\langle 8,4\rangle$ -1367.77 (3.77)	$\langle 2,0\rangle$ vs. $\langle 8,4\rangle$ 898.34 (2.53)	$\langle 1,1\rangle$ vs. $\langle 8,4\rangle$ 789.88 (7.45)
The effect of an additional variable				
$\langle 4,2\rangle$ vs. $\langle 4,3\rangle$	-21.153 (1.186)	-102.482 (1.283)	73.980 (1.215)	70.411 (2.061)
$\langle 4,2\rangle$ vs. $\langle 5,2\rangle$	40.509 (1.426)	-58.020 (1.528)	42.190 (1.394)	37.289 (2.310)
$\langle 8,0\rangle$ vs. $\langle 8,1\rangle$	-17.526 (2.008)	-103.054 (2.701)	66.715 (2.021)	66.391 (3.724)
$\langle 8,0\rangle$ vs. $\langle 9,0\rangle$	5.629 (2.401)	-66.672 (3.416)	44.093 (2.443)	45.414 (4.656)

Cluster corrected standard errors in parentheses. $\langle 8,4\rangle$ notation means a model with 8 discrete and 4 continuous variables, etc. See also Table 2 for additional explanations and primary effects.

nitudes of the effects under consideration gives rather similar results, as far the standard errors of most variables were quite close to each other. This is due to the fact that the simulation design was nearly orthogonal. Most of the settings were used independently of each other, except for the theoretical explained variance dependent on the number of indicators. The settings for the number of categories and the threshold structure were also strongly related to the number of discrete variables, but were randomized in order to achieve the overall balance.

When interpreting those coefficients and their t-statistics, one should keep in mind that only the share of explained variance is a stand-alone regressor, while all others are dummy variables relating the factor of interest to the base. Thus, the lognormal distribution is being compared to the normal distribution of ξ, and the types of analysis are compared to the PCA based on the original continuous variables.

Let us return to the most important findings. The fourth of those, as was noted above, is that the number of variables, and whether those variables are discrete or continuous, plays a key role in performance of PCA. Table 3 shows the magnitudes of those effects. The first line is the largest difference across the estimated coefficients, usually between a model with two indicators, and a model with 12 indicators. The reported figure and corresponding standard error is the difference of the coefficients of the first and the second models mentioned, with notation $\langle p_d, p_c \rangle$ for a data set with p_d discrete and p_c continuous variables. Most of the "Share of explained variance" results are not readily interpretable. In the next column, the linear predictor $x'\hat{\beta}$ from a probit regression for the rank correlation is by 1.368 lower for the model with 2 discrete and no continuous variables than for the model with 8 discrete and 4 continuous variables. This may translate to a difference in the rank correlations as large as 0.60 for the weaker model vs. 0.96 for the model with 12 variables[12]. Likewise, from the differences between the coefficient estimates, the differences between the misclassification rates for different number of variables may be as large as 64% vs. 26% for the overall rate, and 50% vs. 17% for the first quintile.

Table 3 also compares the effects of adding an extra variable. The improvement due to a continuous variable is larger than that for a discrete one by some 60–80%. This can be viewed as a crude measure of the losses to discreteness: roughly speaking, 10 discrete variables contain about as much information, for the PCA purposes, as 6 continuous ones do.

Fifth, there was a number of rather surprising findings. The first one is that the sample size does not matter much, at least for the levels of the dependent variables. The coefficient of the log sample size variable is never larger than 10 in absolute value, which translates to about 2% change in the indicator from the smallest sample size (100) to the largest (10000). To compare, the losses due to the lognormal distribution may be as large as 30% in the Q1 misclassification rate, and due to Filmer-Pritchett procedure, 20% of reported explained variance. Threshold structure had a mixed effect: concentration of half of the observations in a single category usually has a negative effect, but not for the misclassification rates in Q1, where concentration of the observations in the wealthier categories gives more resolution at the left tail of the welfare distribution. The opposite extremes case when the marginal distributions are concentrated on different tails for two different ordinal variables, although known to pose difficulties for the polychoric correlation coefficient estimation, does not have overly detrimental consequences. Likewise, the differences in factor loading that translate to the strength of the relation between the latent ξ and the corresponding indicator, and thus to a greater explanatory power for that variable, although have predictable directions, are not very large in the absolute values.

[12] See Appendix C for interpretation of the values of rank correlations and relations to the misclassification rates.

3.3 Graphical representation

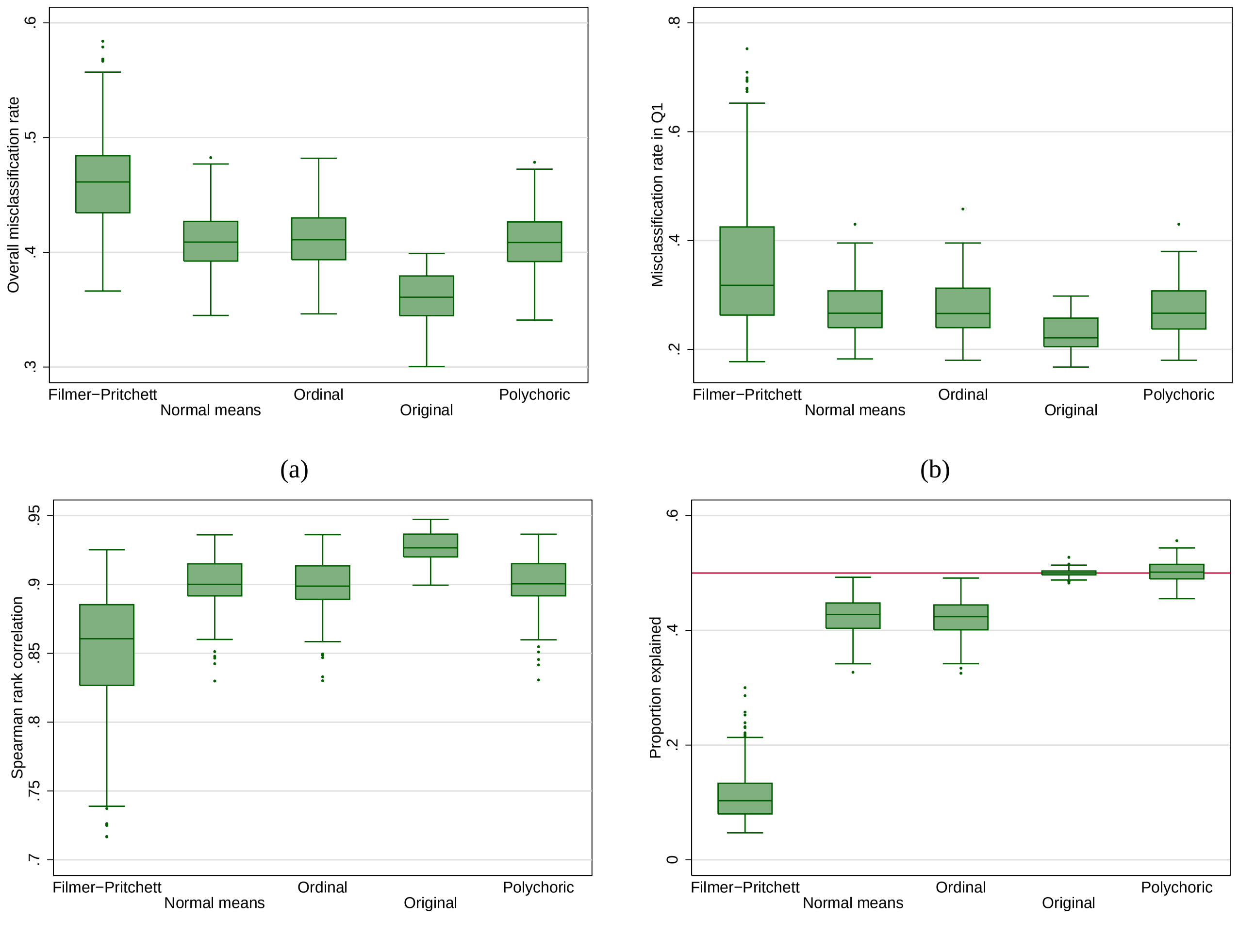
A saying goes that one picture is worth a thousand words. Let us complement the regression analysis with a graphical illustration of our findings. As far as the selection probability weights differ between settings with different numbers of discrete and continuous variables, those differing probabilities of selection would make it hard to come up with clearly interpretable graphs comparing results for different number of variables. We thus confine our attention to a specific setting with 8 discrete and 0 continuous variables, since for this setting, the differences between methods would be quite pronounced due to discrete variables present without any continuous ones. Also, this is the setting with considerably many observations (12880).

The graphical representation is complementary to the results reported in the previous section. The primary value of graphs is, of course, an easy grasp of the distributional features, and we shall draw readers attention to those features in our interpretation of the graphs we report. On one hand, the graphs do not provide any inferential measures such as p-values. But due to the large sample sizes used for each graph, the features detected by eye would probably be interpretable as those of the population. Also, there may be some confounding due to complex simulation design, as even with most involved graphics, it is difficult to visualize more than three or four dimensions, or sources of performance variability, in our case. Due to nearly orthogonal simulation design, we hope the influence of those confounding factors to be minimal. We control for the strongest effects reported in the previous section such as the lognormality of ξ and the underlying proportion of explained variance, so that the graphs really demonstrate the differences in methods along with their sampling variability.

Figure 3 shows the box-and-whisker plots[13] of the four performance indicator discussed in Section 3.2. As in the explanation of Table 2, in the first row of figures, less is better; in picture (c), more is better; and for picture (d), the target is shown with the horizontal line at 0.5.

The best performance is demonstrated by the field-infeasible analysis of the original x^* variables. Also, the Filmer-Pritchett procedure is clearly inferior in all of the analyses. For instance, for the Q1 misclassification rate, the median of the distribution for the Filmer-Pritchett procedure is above the 75-th percentiles of other methods, and the upper quartile of this performance measure is above *all* of the observations for other methods (i.e., in 25% of the worst cases, it does a poorer job than you would ever

[13] 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% of the normal distribution outside the whiskers.



(a)

(b)

(c)

(d)

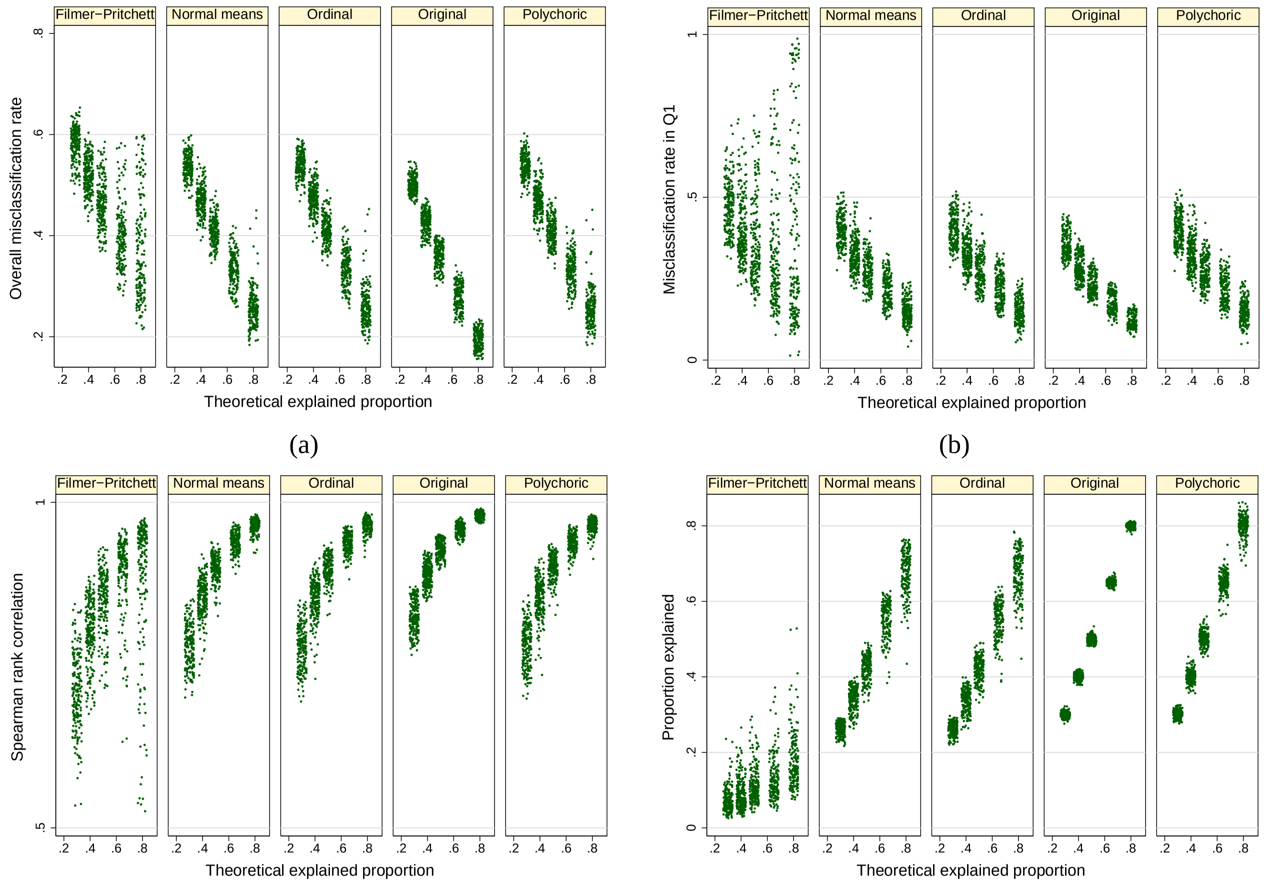
Figure 3: Box plots for different PCA methods. (a) Overall misclassification rate; (b) misclassification rate in the first quintile; (c) Spearman's ρ between the theoretical and empirical welfare measures; (d) share of explained variance. Restrictions: 8 discrete variables, no continuous variables, sample sizes 2000 or 10000, lognormal distribution excluded, theoretical share of explained variance is 0.5.

expect from other methods). Likewise, for the overall quintile misclassification rate, the misclassification rate better than about 43% occurs in about 75% of cases for the ordinal, normal means and polychoric methods, but only for 25% cases for the Filmer-Pritchett method. The other three discrete methods show practically indistinguishable performance, with ordinal PCA giving slightly larger variability, as evidenced by the size of the box.

As for the internal measure of fit, i.e., the reported proportion of explained variance, only the analysis of the original variables and the polychoric PCA show consistency of the reported explained proportion (the graphs are drawn for large sample sizes 2000 and 10000). Other methods are demonstrating the lack of explained variance in the first PC, and the Filmer-Pritchett procedure shows particularly bad bias, with no observations higher than 0.3 even though the target explained variance is 0.5. Based on all four characteristics together, the polychoric method gives the most accurate picture.

As was claimed in Section 3.2, the most important factor in the performance of the PCA, or the most important setting of the simulation study, with the highest t-statistics in Table 2, is the underlying proportion of the explained variance (not to be mixed with the reported proportion of explained variance used as the performance measure!). Fig. 4 shows the relation of our performance measures to the underlying theoretical proportion of the explained variance. The misclassification rates (panels (a) and (b)) show almost linear decline for methods other than Filmer-Pritchett. The latter surprisingly shows increase in variability over the whole range of the explained variances for Q1 misclassification, and for the proportion of explained variance equal to 0.65 and 0.80, for the overall misclassification rate. The reported share of explained variance (panel (d)), although approximately unbiased 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 Filmer-Pritchett procedure demonstrates quite extended lower tail of the distribution that is also very protruded for the proportion of explained variance equal to 0.80.

The next set of findings is related to the number of categories of the discrete variables used in PCA. Those are depicted on Fig. 5. Note that with just two categories (binary indicators like ownership of an asset), the Filmer-Pritchett and ordinal PCA coincide. But as extra categories are added, the performance of the methods does differ notably. For the methods other than Filmer-Pritchett, the four measures approach their “continuous case” limits approximately exponentially and come to saturation at about 5 or 6 categories (except for the proportion of explained variance), consistent



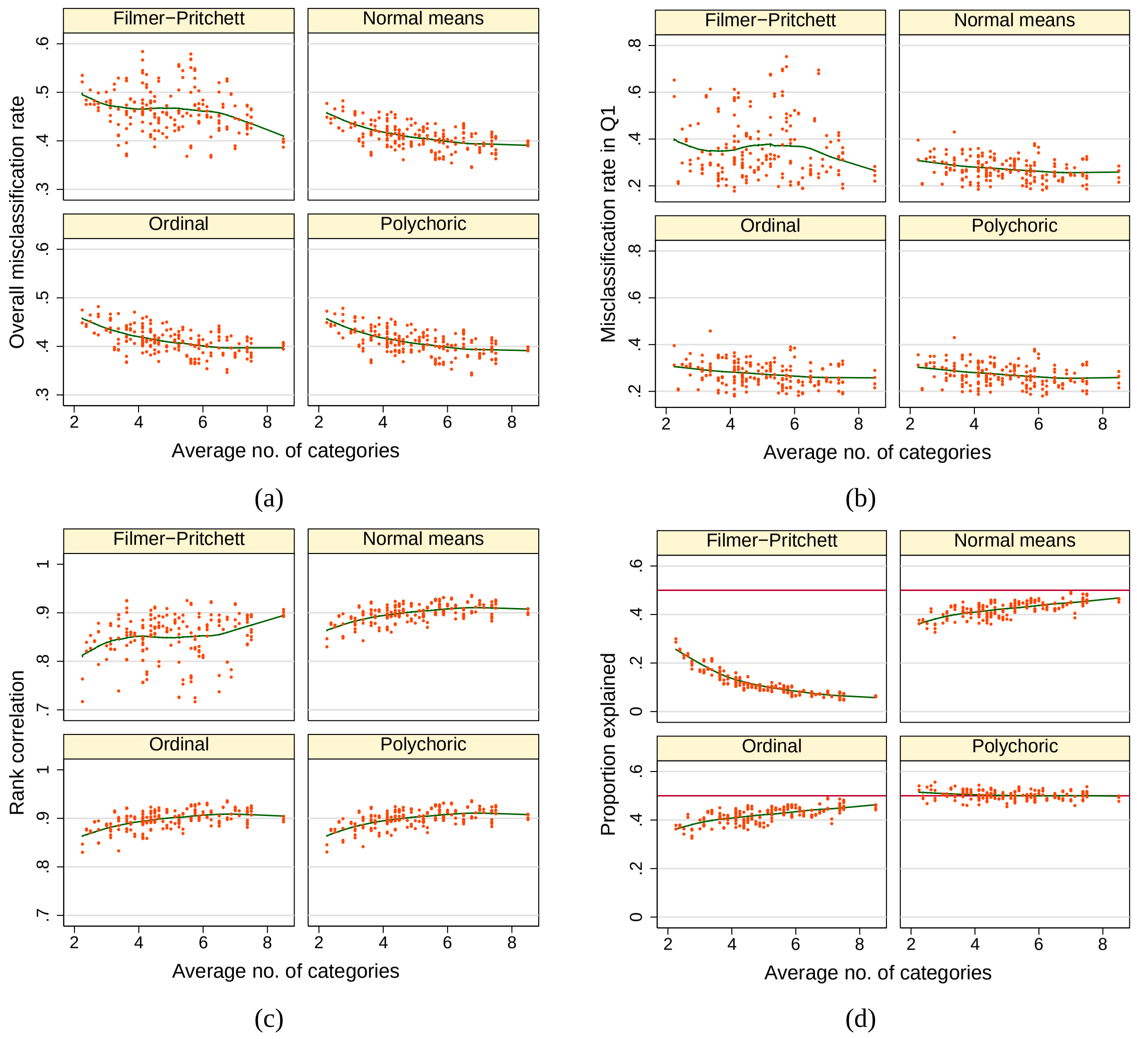
(a)

(b)

(c)

(d)

Figure 4: 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 2000 or 10000, lognormal distribution excluded. Jitter added to show structure.



(a) (b) (c) (d)

Figure 5: 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 2000 or 10000, lognormal distribution excluded, theoretical share of explained variance is 0.5.

with recommendations from the quantitative sociology literature (Dolan 1994). The performance of Filmer-Pritchett procedure also improves with the 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, and the explanation we can propose is that more and more of the spurious and irrelevant negative correlation structure is added to the correlation matrix used as an input to PCA. Also, the amount of information that can be explained by a univariate summary stays about the same as more variables are generated, while the number of variables increases. The former serves as a numerator of (3), and the latter as its denominator. Thus the resulting shape is approximately hyperbolic in the number of categories, which is what is observed in panel (d). See also discussion in Section 2.2. As for the 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.

In all of the above, the observations with lognormal distribution of ξ were excluded, as they led to substantial deterioration of the performance of every method. If those were shown on the graphs in Fig. 3–5, they would look like an extra cloud of points in the direction of deteriorating performance: somewhat below others on the rank correlation and explained variance plots, and somewhat above others, on the misclassification rates plots.

Other combinations of the number of discrete and continuous variables produced qualitatively similar results, although with more continuous variables, the differences between the methods were not as distinct as in the reported case.

4 Conclusion

This paper was motivated by recent examples of use of the principal component analysis in development economics literature starting from Filmer & Pritchett (2001), and investigated several ways to use categorical (in particular, ordinal and binary) variables in the principal component analysis. As far as the distributions of the indicators are non-normal, some of the asymptotic properties of the principal components no longer hold or need to be modified, as the variances and covariances of both eigenvalues and eigenvectors depend on the fourth moments of the data. Other 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 developed several analytical examples demonstrating that (i) categorical variables do have excessive skewness and kurtosis (Example 1); (ii) correlations between categorical variables are on the smallish side (Example 2), (iii) naïve principal component analysis based on the dummy variables aims at placing the two largest groups on the opposite ends of the first principal component score spectrum, and underestimates the proportion of explained variance (Examples 4 and 5).

We then discussed several options that may be useful in performing the principal component analysis in presence of the categorical variables: using ordinal variables per se; using the group means implied by a normal distribution; using the dummy variables for categories as suggested by Filmer & 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. The performance measures used were the quintile misclassification rates (overall and in the first quintile), Spearman rank correlation between the true welfare index used to generate data and the empirical one obtained through the versions of PCA (as an overall measure of the conformance of the rankings of individual observations obtained by the two welfare indices), and the reported proportion of the explained variance, as the main (and often the only one used) measure of the performance available to the researcher.

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 materials used in roofing, dividing the variable into a set of dummy indicators as suggested by Filmer & Pritchett (2001) leads to deterioration of performance according to all of our performance measures used. The explained variance is most heavily affected (underestimated), and more so the more categories are there in the original 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 characteristic of other methods. We thus believe that the researcher will 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 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 in getting the "correct" variable weights may not be very large compared to the PCA on ordinal data. However, only the polychoric analysis gives consistent estimates of the explained proportion. All other methods (PCA on ordinal variables or group means, and even to a greater extent, the Filmer-Pritchett procedure) produce estimates of the proportion of explained variance biased downwards. The misclassification rates, as well as Spearman correlation of the theoretical and empirical welfare indices, are not substantially different between the ordinal, group means and polychoric versions of PCA, although the difference is statistically significant due to huge sample size of the simulation results data set.

Thus, if the researcher uses the Filmer-Pritchett procedure, she should expect the variability of the resulting scores to be higher, and accuracy of the scores to be worse, than those of other methods for discrete PCA we have considered. She would also be very much misled by the reported proportion of explained variance. If the researcher is to choose among ordinal, group means and polychoric PCA, the only reason to prefer one of the methods seems to be the proportion of explained variance, which is reported correctly only by the polychoric method. The rankings, and hence quintile groups, produced by the three methods are very similar to each other.

The performance of PCA also depends on a large number of factors. Expectedly, 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 skewed and heavy tailed distribution of the underlying factor (represented in our case by the lognormal distribution) leads to a notable deterioration of the PCA performance. Note that skewness per se may not be a big problem, as the skewed but not heavy tailed distribution (a mixture of two normals) did not lead to the performance deterioration on the same scale as the heavy tailed distribution. 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, although demonstrated to be crucial in simple settings such as Example 2, were found to be of marginal importance for performance of PCA.

Those results are by and large similar to what is known in the practice of structural equations with latent variables. They also confirm the expectations outlined in simpler settings in the theoretical part of the paper. They also should be viewed in the light of the particular data generating model.

Acknowledgements

Ken Bollen made suggestions that were crucial in the development of the paper, and Nash Herndon provided useful editorial comments. Authors are also grateful for the comments on their poster presentation at the Joint Statistical Meetings (Toronto, 2004). Financial support was provided by U.S. Agency for International Development through MEASURE Evaluation project of Carolina Population Center, University of North Carolina at Chapel Hill, under the terms of Cooperative Agreement #GPO-A-00-03-00003-00.

Appendices

A Practical issues in PCA

The principal component analysis is aimed at solving the (conditional) variance maximization problem (1). This problem turns out to be identical (Anderson 2003, Mardia et al. 1980) to the *eigenvalue* problem (2) that is discussed in Appendix B. Along with the theoretical properties of this linear algebra problem, a researcher is usually interested in statistical properties of this procedure, and in practical uses of its results. This appendix highlights some distributional results available for PCA, and discusses the choice of the number of "significant" components.

The issue of selecting an appropriate model dimensionality does not usually arise in the construction of the welfare index in the household studies, as the first component is the only one that is used, but at least it is worth checking that the first component really stands out relative to the second one, and others. If the first two eigenvalues are relatively close to each other, then the first component may not be very stable, and thus the resulting rankings of the households by their estimated welfare may be misleading.

In other applications such as exploratory data analysis, the researcher often faces the problem of what constitutes a good description of the data in the most concise terms. For the PCA, this is the question of choosing the number of components that the analyst will be using further in her analysis. Most often, this is done graphically by plotting the eigenvalues and eye-balling the place on the graph where the decline in eigenvalues switches from roughly exponential to roughly linear. The plot is referred to as *scree plot*. An example is shown in Fig. 6. The first principal component really stands aside, while the last four or five show a linear trend in eigenvalues. The conclusion from this particular scree plot might be that two or three PCs are "significant" while others represent “noise”.

If a sample from a multivariate normal distribution is taken and PCA is performed on the sample covariance matrix, then the resulting $\hat{\lambda}_i$ and $\hat{\mathbf{v}}_i$ are the maximum likelihood estimates of the corresponding population parameters λ_i, v_i (Mardia et al. 1980). A number of theoretical results on the asymptotic distributions of eigenvalues and eigenvectors can be established under the assumption of normality when the dimension p is fixed and the number of observations $n \to \infty$ (Anderson 1963, Mardia et al. 1980, Theorem 8.3.3):

$$\hat{\lambda}_i \xrightarrow{p} \lambda_i \tag{20}$$

$$\sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow{d} N(0, 2\,\mathrm{diag}(\lambda_i)) \tag{21}$$

$$\sqrt{n}\,\mathrm{Cov}[\hat{\lambda}_i, \hat{v}_j] \to 0\,\forall\, i, j \tag{22}$$

The factor loadings however are not uncorrelated. Also, their variances involve the terms of the form $\lambda_j \lambda_k / (\lambda_j - \lambda_k)$, which are undefined in case of the multiple eigenvalues (and, as we already know, there are no unique eigenvalues, but only a unique eigenspace), and are large for close eigenvalues. The practical implication of this would be the need to check if the second largest eigenvalue is distant enough from the largest one. If it is not, the weights of the variables will be unstable.

Based on those asymptotic results, the likelihood ratio type tests of “significance” of the last $p - k$ components can be constructed (Mardia et al. 1980, section 8.4.3).

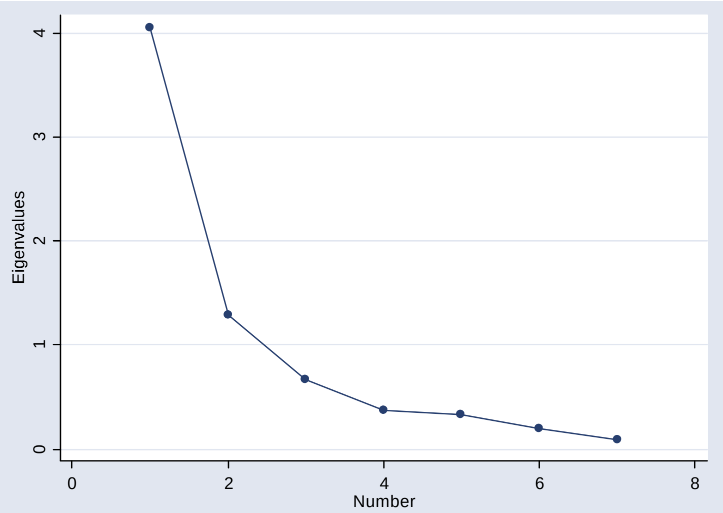


Figure 6: Scree plot. How many components are “significant”?

The null hypothesis is that the first k components have eigenvalues distinctly greater than the remaining $p-k$, and the latter are equal to each other (which is interpreted as having k significant factors, and the rest is the white noise). The test statistic is

$$LR = n(p-k)\ln(a_0/g_0), \tag{23}$$

$$a_0 = \frac{1}{p-k}\sum_{i=1}^{p}\hat{\lambda}_i, \tag{24}$$

$$\ln g_0 = \frac{1}{p-k}\sum_{i=1}^{p}\ln\hat{\lambda}_i, \tag{25}$$

so that a_0 and g_0 are the arithmetic and geometric mean of the eigenvalues that are hypothesized to be equal. The test statistic has an asymptotic χ^2 distribution with $\frac{1}{2}(p-k+2)(p-k-1)$ degrees of freedom. It can be Bartlett corrected by replacing n in (23) with $n-\frac{2p+11}{6}$.

If the distribution of the original data is not normal, Davis (1977) establishes that the asymptotic distributions of the eigenvalues and eigenvectors is still a multivariate normal, but the variances and covariances involve the fourth order cumulants

$$\kappa_{ijkl} = \frac{\partial^4}{\partial t_i \partial t_j \partial t_k \partial t_l}\phi_X(\mathbf{t}) = \mathbb{E}[x_i x_j x_k x_l] - \delta_{ij}(\delta_{ik}\delta_{jl}+\delta_{il}\delta_{jk}) - \delta_i\delta_k\delta_{ij}\delta_{kl} \tag{26}$$

that are identically zero for the normal distribution. (Here, $\phi_X(\mathbf{t}) = \mathbb{E}\big[e^{iX'\mathbf{t}}\big]$ is the characteristic function of the random variable X and its associated distribution.) In particular,

$$\sqrt{n}(\hat{\lambda}_i - \lambda_i) \xrightarrow{p} N(0, 2\lambda_i^2 + \kappa_{iiii}) \tag{27}$$

and it is correlated with other eigenvalues through the fourth order cumulants.

An alternative asymptotics, sometimes referred to as Kolmogorov asymptotics, is to let both the dimension and the number of observations increase in a coherent way so that $n/p \to$ const (Johnstone 2001). Then the consistency of the eigenvalues does not hold any longer, and the spectrum of the estimated eigenvalues is non-degenerate even for the spherical Gaussian distribution with identity covariance matrix.

Additional difficulties can arise due to complex sample design. The researcher needs to make sure that the aspects of it such as weights, clustering, and stratification are accounted for properly. Skinner et al. (1986) compare the results of PCA based on the naïve estimate of the sample covariance matrix as if the data were i.i.d.; model-based estimate that assumes multivariate normality, and accounts for the data known prior to sampling (and used in stratification), and design-unbiased estimator. They show that the first estimator gives biased estimates of both the eigenvalues and eigenvectors when the design calls for weighting, and both the direction and the magnitude of bias depend on the specific designs. The maximum likelihood estimate achieves best results in their simulations, which is not surprising given that they sampled from a multivariate normal distribution. The design-unbiased estimator was unconditionally unbiased, although showed substantial variability, performing well in some samples, and poorly in others (which is a small sample effect, as asymptotically it gives the correct answer). Skinner et al. (1986) provide Taylor series expansion that predicts the deviations (and, eventually, the bias) from the true eigenvalues quite well, but requires the true covariance matrix to be known. Also, the bias depends on the correlation between the stratification variables and the principal component.

The message for the particular applications we are considering here (that of socio-economic status assessment based on DHS data) is that the design does have an effect on the estimates through both (i) design features such as weights, and (ii) through the correlation between the stratification variables (geography) and the substantive first principal component, as there usually is a quite distinct differences in SES levels between regions. The existing software (`polychoric` Stata module) does allow for weights, which is the main source of discrepancies in Skinner et al. (1986).

The geometry of the principal components may be represented graphically in various ways. An obvious way is to use the scores of the observations for the first few components and draw the scatterplots of one principal component against another. For a large data set with more than a hundred or so observations, the picture may start looking messy, so one may consider plotting the centers of reasonably grouped observations.

The usual way to show the *variables* graphically is to plot the factor loadings to show the relation between the principal components and the original variables. This sort of a graph allows clustering by eye of the variables that convey similar information.

B Eigenproblems for real matrices

The general formulation of an eigenproblem for a matrix R with real entries is to find the scalars λ and non-zero vectors v such that

$$R\mathbf{v} = \lambda\mathbf{v}, \quad \|v\| = 1 \tag{28}$$

This is a standard linear algebra problem (Parlett 1980, Horn & Johnson 1990, Weisstein 2004) with applications ranging from acoustics to quantum mechanics, and from statistics to nonlinear optimization. The numbers λ_k are called *eigenvalues*, and the vectors $\mathbf{v}_k$, *eigenvectors*. A number of theoretical properties can be established for them:

1. The eigenvalues are solutions to the *characteristic equation*

   $$\det[R - \lambda I_p] = 0 \tag{29}$$

   It implies that there are p such λ's, although some may repeat, and for an arbitrary matrix R, the eigenvalues may be complex.

2. The eigenvectors $\mathbf{v}_i$, $\mathbf{v}_j$ corresponding to distinct eigenvalues $\lambda_i \neq \lambda_j$ are orthogonal: $\mathbf{v}_i'\mathbf{v}_j = 0$. If λ is a multiple root of equation (29) of order l (such eigenvalues are also referred to as *degenerate*), then there is a linear subspace (an *eigenspace*) of dimension l corresponding to that eigenvalue. Each vector in this subspace satisfies (28) except for normalization, and l orthonormal eigenvectors can be chosen as a basis of that subspace.

3. If $R^T = R$ (i.e., R is real and symmetric), all λ's are real.

4. If R is positive (semi)definite, then all eigenvalues are positive (non-negative).

5. $$\det R = \prod_i \lambda_i, \quad \operatorname{tr} R = \sum_i \lambda_i, \quad \sum_{i,j} r_{ij}^2 = \sum_i \lambda_i^2 \tag{30}$$

6. If R is positive definite, and all off-diagonal entries are non-negative, then the components of the eigenvector corresponding to the largest eigenvalue are all positive.

Most of the time, the eigenvectors are taken to have unit length for identification, and the eigenvalues are ordered from the largest to the smallest: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. The set $\{\lambda_k\}$ is also referred to as the *spectrum* of the matrix R.

Some numeric linear algebra considerations (Demmel 1997) should be taken into account in the applied analysis. The most important one is how well is the matrix conditioned. The ratio λ_1/λ_p is referred to as the *condition number*[14] of the matrix. The condition number is the (upper bound of the) multiplier for the relative error of a linear algebra algorithm (such as solution of a system of linear equations, matrix inversion, or eigenproblem). In other words, it shows by how much the relative error can be expected to go up from the inputs of the algorithm to its output. The relative error in double precision arithmetics is about 10^{-15}, so the condition number of a

[14] With respect to the operator, or spectral, norm of the matrix.

matrix of the order 10^{15} means that the linear algebra problems cannot be solved in double precision for this matrix. Ill-conditioned matrices should thus be avoided, and an obvious example of such a matrix is the covariance or correlation matrix of the dummy variables that sum up to one (i.e., no category was taken to be the base and excluded). For such a matrix, the condition number is infinity. In practice, the solution of the eigenproblem may still yield small non-zero eigenvalues due to round-off errors, but the condition number would still be very high signalling the problem. Round-off errors can also make some of the zero eigenvalues of a positive semi-definite matrix negative, which is another signal for numeric problems in PCA. The modern software is likely to have ways around such problems such as early automatic detection of zero eigenvalues[15], but the researcher should not completely rely on this, and make the computations more efficient by unlinking the dummy variables from each other.

From the statistical point of view, the condition number can be viewed as a crude measure of dependence between standardized variables if the PCA is performed on the correlation matrix: if the variables are independent, all eigenvalues would be equal to 1, and the condition number is 1. For the dependent data, the condition number will be greater than 1. It may also be useful for collinearity diagnostics in the regression context.

The following example shows the relation between the eigenvalues of two correlation matrices with attenuated correlations.

Example 3. If C_1 and C_2 are correlation matrices of the same size such that their off-diagonal entries are proportional to each other:

$$c_{ii}^{(1)} = c_{ii}^{(2)} = 1, c_{ij}^{(1)} = \alpha c_{ij}^{(2)}, 1 \le i \ne j \le p \tag{31}$$

then C_1 can be represented as a convex combination:

$$C_1 = \alpha C_2 + (1-\alpha) I_p \tag{32}$$

and the following holds:

$$\begin{aligned} 0 = \det(C_1 - \lambda^{(1)} I_p) &= \det(\alpha C_2 + (1-\alpha)I_p) - \lambda^{(1)} I_p) = \\ &= \det\big(\alpha[C_2 - \lambda^{(1)}]\big) = \det(C_2 - \frac{\lambda^{(1)} - 1 + \alpha}{\alpha} I_p) \end{aligned} \tag{33}$$

so the eigenvalues of the two matrices are related as follows:

$$\lambda^{(1)} - 1 = \alpha(\lambda^{(2)} - 1) \tag{34}$$

[15] This is what Stata software does with collinearity diagnostics by dropping some of the variables it thinks are responsible for it.

In other words, the spectrum of C_1 is shrunk towards 1 relative to C_2. In particular, the largest eigenvalue of C_1 is greater than one, but not by as much as the largest eigenvalue of C_2. So in the PCA on the two matrices, the first PC will explain more variance for the matrix C_2 than for C_1.

If there is no strict proportionality between the entries of two matrix, a similar argument can still be made by the Gershgorin circles theorem and its extensions (Brualdi & Mellendorf 1994). The convex combination (32) would become

$$C_1 = \alpha C_2 + (1-\alpha)A \tag{35}$$

where α and A are found from

$$a_{ii} = 1, \quad \alpha = \arg\min_{\alpha} R_\alpha, \quad R_\alpha = \max |a_{ij}|, i,j = 1, \ldots, n \tag{36}$$

For the original problem, R_α can be brought down to zero. Then the relation (34) between eigenvalues of the two matrices would also need to be attenuated by rather crude upper bounds R_α of the absolute value of the off-diagonale elements. ⊠

C Rank correlations

A standard measure of the relation between the two variables is their correlation coefficient:

$$\rho = \mathbb{E}\,\frac{(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])}{(\mathbb{V}[X]\,\mathbb{V}[Y])^{\frac{1}{2}}} \tag{37}$$

(also referred to as Pearson moment correlation), and its sample analogue

$$r = \frac{1}{n-1}\sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x}\,\frac{y_i - \bar{y}}{s_y} \tag{38}$$

Rank correlations deal with ranks r_x and r_y given to the observations by two variables x and y, rather than the values (x_i, y_i) themselves. Hence the rank correlations are invariant under any monotone transformation of the original variables, not only under linear transformations, as is the moment correlation. Thus the rank correlation of say acres of land owned with income will be the same as the rank correlation of the acreage with log income. Also this is more useful in our analysis if the underlying welfare ξ and its empirical analogue produced by a version of PCA have a curvilinear rather than a linear relation. In the latter case, the two will produce the same distribution quintiles, even though the moment correlation will signal the departure of one from another. Finally, our interest in the rank correlations is due to the use of PCA in ranking observations (households) according to their estimated welfare into quintiles, as in Filmer & Pritchett (2001) or Demographic and Health Survey reports.

The most frequently used version of a rank correlation is *Spearman rank correlation*, which simply uses the formula (38) with x and y replaced by their implied ranks r_x and r_y. When there are no ties, it can also be written as

$$\rho_S = 1 - \frac{6}{n^2(n-1)} \sum_{i=1}^{n} (r_{x,i} - r_{y,i})^2 \tag{39}$$

Another version of rank correlation is referred to as *Kendall's* τ. If C is the number of concordant pairs of observations (i.e., both x and y place one of the observations higher than the other), D is the number of discordant pairs, the total number of pairs is $N = n(n-1)/2$, then the two versions of Kendall's correlation are defined as

$$\begin{gathered} S = C - D, \qquad \tau_a = S/n, \\ \tau_b = \frac{S}{\sqrt{(n-U)(n-V)}}, \\ U = \sum_i u_i(u_i - 1)/2, \quad V = \sum_i v_i(v_i - 1)/2 \end{gathered} \tag{40}$$

where u_i is the multiplicity of the value x_i, and v_i is the mutliplicity of the value y_i. Thus τ_b corrects for the ties in the data set.

The interpretation of Kendall's τ is the (relative) number of pairwise transpositions one would need to make to reorder the data so that two variables agree; and, as is clearly seen from the definition, the proportion of concordant vs. discordant pairs of observations.

All the aforementioned correlations can be embedded into the following general formula:

$$\rho(d) = \frac{\sum_{i,j} d(x_i, x_j) d(y_i, y_j)}{\left[\sum_{i,j} d^2(x_i, x_j) \sum_{i,j} d^2(y_i, y_j)\right]^{1/2}} \tag{41}$$

where the generalized measure of discrepancy $d(\cdot)$ between the two observations i, j as given by variables x and/or y is:

- $d(x_i, x_j) = \mathrm{sign}(x_i - x_j)$ for Kendall's correlation ($+1$ if the rankings agree, and -1 if they disagree);
- the difference of the ranks $d(x_i, x_j) = r_{x,i} - r_{x,j}$ for Spearman's ρ;
- $d(x_i, x_j) = x_i - x_j$ for Pearson moment correlation.

The distributions of both rank correlations under the null hypothesis of independence can be derived by noting that all $n!$ combinations are equally likely, and then counting the number of combinations that give a particular value of ρ_S or τ. Those are examples of discrete distributions even though the values of the random variable are not integers. The asymptotic distribution of either quantity is normal since it is based on the sums of identically distributed random variables. In particular, those correlation coefficients are combinations of U-statistics which are known to be asymptotically normal (Hoeffding 1948, van der Vaart 1998).

It follows from (41) that Kendall's τ gives smaller weights to the pairs of observations that have drastically different ranks than Spearman's ρ_S. Kendall (1955) shows that in large samples, $-1 \leq 3\tau - 2\rho_S \leq 1$, although the limits are attained for rather peculiar rankings. Another limiting inequality for large samples and $\tau > 0$ is $\frac{3}{2}\tau - \frac{1}{2} \leq \rho_S \leq \frac{1}{2} + \tau - \frac{1}{2}\tau^2$. Those inequalities show that usually τ tends to be smaller in magnitude than ρ_S, and it is possible that the two rank correlations come up to be of different signs.

The rank correlations can be linked to more easily understandable quintile misclas-

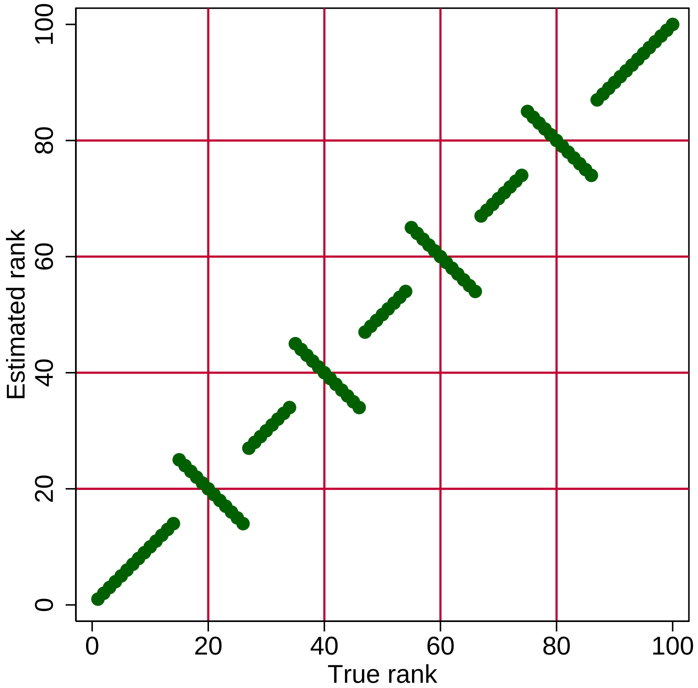


Figure 7: The worst case misclassification rate with the highest Kendall's τ.

sification rates. The worst case scenario is depicted in Fig. 7. In this case, the overall misclassification is the highest, although the Kendall's τ is not that badly affected. If a fraction $0 \leq \nu \leq \frac{1}{2}$ of observations in the bottom of each quintile are wrongly attributed to the previous quintile, and the same number of observations from the top of the quintile are attributed to the next quintile, then the overall misclassification rate is 1.6ν (so at the maximum when $\nu = \frac{1}{2}$, the misclassification rate is 80%: only the 10% at the very bottom and 10% at the very top are correctly classified), while the rank correlation is $1 - 1.28\nu^2$ (and even in the worst case it still is quite high at 0.68). Then the upper bound on the misclassification rate (i.e. the worst case relation between the two) is that the misclassification rate is no worse than $\sqrt{2-2\tau}$.

With the above relation between ρ_S and τ, the boundary on the misclassification rate can also be expressed in terms of Spearman's correlation coefficient and becomes $[8(1-\rho_S)]^{1/4}$. This is a very pessimistic bound that gives the misclassification rate of 1 for $\rho_S = 7/8 = 0.875$, and misclassification of 50%, for $\rho_S = 0.99$.

Computation of Spearman's correlation ρ_S requires two additional sortings of the data which is a $O(n \ln n)$ operation. Computation of Kendall's τ is of combinatorial complexity $O(n^2)$ and not practically feasible for samples larger than few hundred, at least not for the simulation purposes.

As a final comparison point, the statistical properties of Kendall's τ are somewhat better studied in the statistical literature. In particular, it achieves asymptotic normality faster than Spearman's ρ_S, so the tests based on τ are more accurate in samples as small as $n = 20$.

D PCA on binary indicators of a single factor

This appendix considers the results of the principal component analysis performed on a set of dummy variables obtained as indicators of categories of another variable.

Suppose there is a single factor ξ, a single indicator, and a categorical version of that indicator with K categories. This would be the case if one uses the Filmer-Pritchett procedure to obtain weights if only a single categorical variable is observed, or it can be the case when many categorical variables are coded in such a way that each unique combination is represented by a corresponding binary indicator of that combination.

The categorical variable in this example will have a multinomial distribution[16]. If

[16] A categorical variable x is said to have a multinomial distribution with categories $1, \dots, K$ and probabilities $p_1, \dots, p_K$ if for the sample of size n,

$$\text{Prob}[|\{i : x_i = 1\}| = n_1, \dots, |\{i : x_i = K\}| = n_K] = \frac{n!}{n_1! \dots n_K!} p_1^{n_1} \cdot \dots \cdot p_K^{n_K} \qquad (42)$$

This is a natural generalization of the binomial distribution into more than two categories. See Johnson, Kotz the proportion of the data in k-th category is π_k, then the covariance matrix of the dummy variables corresponding to the individual categories is

$$\Sigma = \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & \multicolumn{2}{c}{\dots\dots\dots\dots\dots\dots} \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) & \multicolumn{2}{c}{\dots\dots\dots\dots\dots\dots} \\ \vdots & \vdots & \ddots & \vdots \\ \multicolumn{2}{c}{\dots\dots\dots\dots\dots\dots} & -\pi_{K-1,K} & \pi_K(1-\pi_K) \end{pmatrix} \qquad (43)$$

The next example deals with the case when there is the same number of observations in each category.

Example 4. In a special case when all categories are equally populated ($\pi_1 = \pi_2 = \ldots = \pi_K = 1/K$), the diagonal elements are $(K-1)/K^2$, and off-diagonal elements are $-1/K^2$. The correlation matrix is then

$$\Sigma' = \begin{pmatrix} 1 & -\rho & -\rho & \dots \\ -\rho & 1 & -\rho & \dots \\ \vdots & \vdots & \ddots & \vdots \\ \dots & -\rho & -\rho & 1 \end{pmatrix} \qquad (44)$$

where

$$\rho = \frac{1/K^2}{(K-1)/K^2} = \frac{1}{K-1} \qquad (45)$$

Then by using the symmetry arguments and direct verification[17], the eigenvalues and the corresponding eigenvectors are zero with an eigenvector[18] of $K^{-\frac{1}{2}}(1,\ldots,1)$, and $K-1$ eigenvalues of $K/(K-1)$ with an eigenspace[19] generated by the vectors of the form $u_i = (1,\ldots,0,-1,0,\ldots,0)$ that have -1 in their i-th position. The proportion explained by the first principal component (and, in fact, all non-trivial components) will be $1/(K-1)$. The first principal component is not well defined in this case: any weights that sum up to zero can be taken as the weights for the first PC. As a result, the sample first PC will be extremely unstable. (The problem may be diagnosed in the way described in Appendix A: the analyst would need to have a look at the second component as well, and would find that the first two sample eigenvalues are close to each other.

& Balakrishnan (1997), Chapter 35.

[17] See Appendix E, p. 52.

[18] It corresponds to the bookkeeping condition $\pi_1 + \ldots + \pi_K = 1$. The variance of this condition is identically zero, and that is shown by the zero eigenvalue. If the dummy variables are coded in such a way that some of them sum up to 1, then one or more eigenvalues would be equal to zero. From the theoretical point of view, this does not constitute a significant problem, as the eigenvalues and the scores would be the same should one of the categories be dropped. The zero eigenvalues, however, may be a problem for numerical stability of the eigenproblem algorithms as explained in Appendix B.

[19] See Appendix A, p. 41 on explanation of eigenvectors and eigenspaces.

This result can also be derived at from the asymptotic distributions viewpoint described also in Appendix A.) $\boxtimes$

In the general case, the proportions of categories will be different, and the next example gives a basic analysis of this case.

Example 5. Let us now consider a more realistic setting where $\pi_1 > \pi_2 > \ldots > \pi_K$ (the ordering is assumed for the sake of transparency of the analysis). This will be the general case for discrete data if we simply consider all possible categories together, and create dummy variables for each of them. If there was a natural ordering of the categories, it is disregarded in this analysis.

Now, the correlation matrix becomes

$$\Sigma = \begin{pmatrix} 1 & -\rho_{12} & -\rho_{13} & \ldots \\ -\rho_{12} & 1 & -\rho_{23} & \ldots \\ \vdots & \vdots & \vdots & \ddots \\ \ldots & -\rho_{K-2,K} & -\rho_{K,K-1} & 1 \end{pmatrix} \tag{46}$$

Within each row or column, the values of ρ_{ij} are decreasing, in absolute value, as one moves further away from the diagonal:

$$\rho_{ij} = \frac{\pi_i \pi_j}{\sqrt{\pi_i(1-\pi_i)\pi_j(1-\pi_j)}} = \sqrt{\frac{\pi_i \pi_j}{(1-\pi_i)(1-\pi_j)}} \tag{47}$$

is a monotone increasing function of either π_i or π_j. The entry ρ_{12} is the largest off-diagonale entry, so the principal component analysis / eigenproblem of the matrix (46) will pick up on this entry, assigning the two largest weights of different sign to the indicators of the first and second categories, thus increasing the variance of their linear combination.

In the simplest analytically tractable case, $K = 3$, suppose that the population fractions are $1/3 + \Delta$, 1/3, and $1/3 - \Delta$, where $\Delta \ll 1$ is a perturbation (Judd 1998, chapter 13) — say due to sampling fluctuations. Then by taking the first order expansion with respect to Δ, one can show that the correlation matrix of the perturbed distribution is

$$\begin{pmatrix} 1 & -\frac{1}{2} - \frac{3}{8}\Delta & -\frac{1}{2} \\ -\frac{1}{2} - \frac{3}{8}\Delta & 1 & -\frac{1}{2} + \frac{3}{8}\Delta \\ -\frac{1}{2} & -\frac{1}{2} + \frac{3}{8}\Delta & 1 \end{pmatrix} \tag{48}$$

By solving the eigenproblem for this matrix, one finds that the double eigenvalue of 3/2 splits into $3/2+\sqrt{3}\Delta/4$ and $3/2-\sqrt{3}\Delta/4$ with the zero order terms of eigenvectors proportional to $(1, 1-\sqrt{3}, \sqrt{3}-2)$ and $(\sqrt{3}-2, 1-\sqrt{3}, 1)$ (the normalized versions: $(0.789, -0.577, -0.211)$ and

$(-0.211, -0.577, 0.789)$, respectively)[20]. The third eigenvalue is still identically zero which reflects the fact that the sum of the dummy variables related to a single factor is 1, so that the covariance and correlation matrices are singular. The null space eigenvector is $1/\sqrt{3}(1 + \Delta/4, 1, 1 - \Delta/4)$. The proportion of the variance explained by the first principal component goes up from a half to $1/2 + \sqrt{3}\Delta/4$. ⊠

If we interpret the perturbation of the correlation matrix as the effect of sampling fluctuations under the true distribution that puts equal masses to all of the three categories (i.e., if we analyze the empirical correlation matrix that has a form (48) while the true population correlations are $-1/2$)[21], then the first principal component is highly unstable and swings discretely into an arbitrary combination of the weights approximately equal to ± 0.789, ∓ 0.577, ∓ 0.211 given to the three categories depending exclusively on the sampling fluctuations, i.e., which of the categories happened to have a larger representation in the particular sample.

If the categories were ordered, then only with probability of 1/6 does the first PC give the right ordering of the categories (when the top category is the most populated one, and the bottom, the second largest), and with probability 1/6, the reverse ordering of categories. (In the latter case, the analyst would still be able to recover the "right" ordering if she checks the direction of her empirical first PC.) In the remaining 2/3 of cases, the first PC score puts one of the extreme categories in the middle of the distribution thus breaking the natural ordering. With more categories, the situation is likely to get worse. The practical implication is that if the dummy variables are to be used per se in the PCA, then for stability of the analysis it is desirable that the proportions of the data points in those categories are different enough, and the top and bottom groups are the two largest. If those conditions are violated, the results of the PCA may not make much sense. Of course, in practice one would use several different factors, so the situation may not be that dim if several variables were used to construct the binary indicators.

[20] The first order analysis of the perturbed correlation matrix cannot give the corrections to the eigenvectors a. If the linear system $(\Sigma - \lambda I)a = 0$ is perturbed (Σ is changed into $\Sigma + \Delta\Sigma$), then the first order approximation for Δa is $(\Sigma - \lambda I)\Delta a = -(\Delta\Sigma - \Delta\lambda I)a$. The matrix in the left hand side, however, is not invertible, so there is no unique solution for Δa. Higher order expansions lead to nonlinear matrix problems. The statistical implication of that is high sampling variability of the empirical eigenvectors that really is found in practice: the observed eigenvector may be quite far from $\pm(0.789, -0.577, -0.211)$ even for fairly small deviations from Σ.

[21] The fact that the middle category is not perturbed is not particularly important: the main issue is that the three categories are not equally populated.

E Proofs, derivations and examples

E.1 Applicability of PCA for (5)

Suppose the (continuous, fully observed) data $x_1, \ldots, x_p$ come from the model with one latent variable ξ (c.f. (5)):

$$x_k = \Lambda_k \xi + \delta_k, \tag{49}$$

$$\mathbb{V}[\delta] = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_k^2), \quad \mathbb{V}[\xi] = \phi \tag{50}$$

The simulations in Section 3 make further simplifying assumptions:

$$\begin{gathered}\sigma_k^2 = \sigma^2 \, \forall k \\ \Lambda_k = b, k = 1, ..., p_1; \quad \Lambda_k = 1, k = p_1 + 1, \ldots, p_1 + p_2 = p \end{gathered} \tag{51}$$

Then the covariance matrix of x has the block form

$$\mathbb{V}[x] = \begin{pmatrix} b^2\phi + \sigma^2 & b^2\phi & \ldots & b\phi & \ldots & b\phi \\ b^2\phi & b^2\phi + \sigma^2 & \ldots & b\phi & \ldots & b\phi \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ b\phi & b\phi & \ldots & \phi + \sigma^2 & \ldots & \phi \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ b\phi & b\phi & \ldots & \phi & \ldots & \phi + \sigma^2 \end{pmatrix} \tag{52}$$

and the correlation matrix has a similar structure

$$C = \operatorname{Corr}[x] = \begin{pmatrix} 1 & u & \ldots & w & \ldots & w \\ u & 1 & \ldots & w & \ldots & w \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ w & w & \ldots & 1 & \ldots & v \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ w & w & \ldots & v & \ldots & 1 \end{pmatrix}$$

$$u = \frac{b^2\phi}{b^2\phi + \sigma^2}, \quad v = \frac{\phi}{\phi + \sigma^2}, \quad w = \frac{b\phi}{\sqrt{(b^2\phi + \sigma^2)(\phi + \sigma^2)}} = \sqrt{uv} \tag{53}$$

assuming all correlations to be positive, which will be provided $b > 0$.

E.2 Optimal prediction $\hat{\xi}$

Based on the observed $x_1, \ldots, x_p$, let us construct a prediction of ξ as

$$\hat{\xi} = \sum_k a_k x_k \tag{54}$$

Although it may look like a regression-type prediction, it really is not. The model (50) defines p equations with x_k being the dependent variable, and ξ being the only independent variable in the regression. Rather, this is a problem of *inverse regression*: given the values of the dependent variable(s), construct the best estimate of the explanatory variable.

If $x_1, \ldots, x_p, \xi$ have a multivariate normal distribution (as would be the case if $\xi \sim N(0, \phi)$ and $\delta_k \sim N(0, \sigma^2)$):

$$x_1, \ldots, x_p, \xi \sim N\left(\mathbf{0}, \begin{pmatrix} \Sigma & \tau^T \\ \tau & \phi \end{pmatrix}\right) \tag{55}$$

where $\Sigma = \mathbb{V}[x]$, $\tau = \mathrm{Cov}[x, \xi]$, then by the properties of the normal distribution (Mardia et al. 1980),

$$\xi | x = \mathbb{E}[\xi] + \tau^T \Sigma^{-1}(x - \mathbb{E}[x]) = \tau^T \Sigma^{-1} x \tag{56}$$

which indeed defines a linear combination of x's.

The weights a_k in the linear combination (54) can be found through the matrix formulae in (56), where the inverse matrix Σ^{-1} can be shown to have the same block structure as Σ does, but an easier and more straightforward way can be to note that $\hat{\xi}$ is a projection of ξ onto the space of x's, and as a projection, it should minimize the norm of the squared deviations:

$$\mathbb{E}[\hat{\xi} - \xi]^2 \to \min_{a_1, \ldots, a_k} \tag{57}$$

Note further that the first p_1 variables are the same in their statistical properties, and the permutation of those would not change the covariance matrix in (55), so neither would it change the weights resulting from (56). Thus, the first p_1 values $a_1, \ldots, a_{p_1}$ are identical: $a_1 = \ldots = a_{p_1}$. Likewise, the last p_2 entries are also equal to each other: $a_{p_1+1} = \ldots = a_p$. Let us denote $a_1 = \alpha$, $a_p = \beta$. Then the projection problem becomes

$$\mathbb{E}[\sum_{k=1}^{p_1} \alpha x_k + \sum_{k=p_1+1}^{p_1+p_2} \beta x_k - \xi]^2 \to \min_{\alpha, \beta} \tag{58}$$

Then

$$\begin{aligned} \mathbb{E}[\sum_{k=1}^{p_1} \alpha x_k + \sum_{k=p_1+1}^{p_1+p_2} \beta x_k - \xi]^2 &= \mathbb{E}[\sum_{k=1}^{p_1} \alpha(b\xi + \delta_k) + \sum_{k=p_1+1}^{p_1+p_2} \beta(\xi + \delta_k) - \xi]^2 = \\ &= \mathbb{E}[\sum_{k=1}^{p_1} \alpha\delta_k + \sum_{k=p_1+1}^{p_1+p_2} \beta\delta_k + p_1 b\alpha\xi + p_2\beta\xi - \xi]^2 = \\ &= \alpha^2 p_1 \sigma^2 + \beta^2 p_2 \sigma^2 + (p_1 b\alpha + p_2\beta - 1)^2 \phi \end{aligned} \tag{59}$$

Introducing $\nu = \sigma^2/\phi$, we can rewrite the minimization problem as

$$V(\alpha,\beta) = \alpha^2 p_1 \nu + \beta^2 p_2 \nu + (p_1 b\alpha + p_2\beta - 1)^2 \to \min_{\alpha,\beta} \tag{60}$$

Taking the derivatives yields:

$$\begin{aligned}\frac{\partial V}{\partial \alpha} &= 2\alpha p_1 \nu + 2p_1 b(p_1 b\alpha + p_2\beta - 1) = 0 \\ \frac{\partial V}{\partial \beta} &= 2\beta p_2 \nu + 2p_2(p_1 b\alpha + p_2\beta - 1) = 0 \end{aligned} \tag{61}$$

Solving this gives

$$\alpha = b\beta = b\frac{1}{p_1 b^2 + p_2 + \nu}, \tag{62}$$

$$\hat{\xi} = \frac{1}{p_1 b^2 + p_2 + \nu}\Big(\sum_{k=1}^{p_1} bx_k + \sum_{k=p_1+1}^{p} x_k\Big) \tag{63}$$

The most important implication is that the ratio of the weights in (63) is equal to b, the ratio of the original factor loadings in (50) and (51).

E.3 PCA for (54)

Let us first consider a somewhat simpler case of the eigenproblem for a $p \times p$ matrix

$$\Omega(u) = \begin{pmatrix} 1 & u & \dots & u \\ u & 1 & \dots & u \\ \vdots & \vdots & \ddots & \vdots \\ \dots\dots & u & 1 & \end{pmatrix} \tag{64}$$

If $e_i = (0,\dots,0,1,0,\dots,0)$ is a unit vector with 1 at i-th position, then

$$\Omega(e_1 - e_i) = \begin{pmatrix} 1 \\ \vdots \\ u \\ u \\ u \\ \vdots \end{pmatrix} - \begin{pmatrix} u \\ \vdots \\ u \\ 1 \\ u \\ \vdots \end{pmatrix} = \begin{pmatrix} 1-u \\ \vdots \\ 0 \\ u-1 \\ 0 \\ \vdots \end{pmatrix} = (1-u)(e_1 - e_i) \tag{65}$$

so $1-u$ is a multiple eigenvalue of the order $p-1$ with the corresponding eigenvector $e_1 - e_i$. An orthonormal basis of this eigenspace can be found as the set of vectors

$$u_i = [-1 + (p-1)(p-\sqrt{p})]e_1 + e_i + (p-\sqrt{p})\sum_{j=2}^{p} e_j, \quad i = 2,\dots,p \tag{66}$$

The last eigenvalue can obtained by noting that

$$\Omega \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} 1+(p-1)u \\ 1+(p-1)u \\ \vdots \\ 1+(p-1)u \end{pmatrix} = [1+(p-1)u] \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{67}$$

so the eigenvalue is $1+(p-1)u$ and the corresponding eigenvector is $(1,\dots,1)^T$, with a normalized version $p^{-1/2}(1,\dots,1)^T$.

For a more interesting case (53), let us find the largest eigenvalue (assuming all correlations to be positive) and the corresponding eigenvector:

$$a^t C a \to \max_{a:\|a\|=1} \tag{68}$$

By the symmetry argument similar to the one in the preceding section, the first p_1 elements of a should be identical, and the also last p_2 elements of a should be identical. By denoting them ζ and η, respectively, the quadratic form becomes

$$\begin{pmatrix} \zeta \\ \zeta \\ \vdots \\ \eta \\ \vdots \\ \eta \end{pmatrix}^T \begin{pmatrix} 1 & u & \dots & w & \dots & w \\ u & 1 & \dots & w & \dots & w \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ w & w & \dots & 1 & \dots & v \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ w & w & \dots & v & \dots & 1 \end{pmatrix} \begin{pmatrix} \zeta \\ \zeta \\ \vdots \\ \eta \\ \vdots \\ \eta \end{pmatrix} = \begin{pmatrix} \zeta+(p_1-1)u\zeta+p_2w\eta \\ \zeta+(p_1-1)u\zeta+p_2w\eta \\ \vdots \\ p_1w\zeta+\eta+(p_2-1)\eta \\ \vdots \\ p_1w\zeta+\eta+(p_2-1)\eta \end{pmatrix}^T \begin{pmatrix} \zeta \\ \zeta \\ \vdots \\ \eta \\ \vdots \\ \eta \end{pmatrix} =$$

$$= p_1^2u\zeta^2 + 2p_1p_2w\zeta\eta + p_2^2v\eta^2 = (p_1\sqrt{u}\zeta + p_2\sqrt{v}\eta)^2 \tag{69}$$

The norm of the vector is then $p_1\zeta^2 + p_2\eta^2 = 1$, and the maximization problem is

$$(p_1\sqrt{u}\zeta + p_2\sqrt{v}\eta)^2 \to \max_{\zeta,\eta:p_1\zeta^2+p_2\eta^2=1} \tag{70}$$

The Lagrangian is

$$L(\zeta,\eta,\Lambda) = (p_1\sqrt{u}\zeta + p_2\sqrt{v}\eta)^2 - \Lambda(p_1\zeta^2 + p_2\eta^2 - 1) \tag{71}$$

and the partial derivatives are

$$\begin{aligned} \frac{\partial L}{\partial \zeta} &= 2p_1\sqrt{u}(p_1\sqrt{u}\zeta + p_2\sqrt{v}\eta) - 2\Lambda p_1\zeta = 0 \\ \frac{\partial L}{\partial \eta} &= 2p_2\sqrt{v}(p_1\sqrt{u}\zeta + p_2\sqrt{v}\eta) - 2\Lambda p_2\eta = 0 \end{aligned} \tag{72}$$

so the the weights are

$$\frac{\zeta}{\eta} = \sqrt{\frac{u}{v}} = b\sqrt{\frac{\phi+\sigma^2}{b^2\phi+\sigma^2}} = b\sqrt{\frac{1+\nu}{b^2+\nu}} \approx b\Big(1-\frac{b-1}{1+\nu}\Big) \text{ for } b \approx 1,$$
$$\zeta = \sqrt{\frac{u}{p_1 u + p_2 v}}, \quad \eta = \sqrt{\frac{v}{p_1 u + p_2 v}}, \tag{73}$$

Thus the weights differ from being proportional to $(b, \dots, b, 1, \dots, 1)$, although not very greatly if b is close to 1. The value of the quadratic form is the largest eigenvalue:

$$(p_1\sqrt{u}\zeta + p_2\sqrt{v}\eta)^2 = \left(\frac{p_1 u}{\sqrt{p_1 u + p_2 v}} + \frac{p_2 v}{\sqrt{p_1 u + p_2 v}}\right)^2 = p_1 u + p_2 v \tag{74}$$

so the proportion of the explained variance is

$$R_1 = \frac{p_1 u + p_2 v}{p} = \frac{p_1 u + p_2 v}{p_1 + p_2} \tag{75}$$

If the PCA is performed on the original variables (or their covariance matrix), then we can show that the first principal component will be the same as $\hat{\xi}$ from (63). Indeed, upon invoking the symmetry argument once again, the problem now becomes

$$\mathbb{V}[\sum_{k=1}^{p_1} \zeta x_k + \sum_{k=p_1+1}^{p_1+p_2} \eta x_k] \to \max_{\zeta,\eta: p_1\zeta^2 + p_2\eta^2 = 1} \tag{76}$$

Then the variance of the linear combination in question is

$$\mathbb{V}[\sum_{k=1}^{p_1} \zeta x_k + \sum_{k=p_1+1}^{p_1+p_2} \eta x_k] = \mathbb{V}[\sum_{k=1}^{p_1} \zeta(b\xi + \delta_k) + \sum_{k=p_1+1}^{p_1+p_2} \eta(\xi + \delta_k)] =$$
$$= \mathbb{V}[(p_1 b\zeta + p_2\eta)\xi + \sum_{k=1}^{p_1} \zeta\delta_k + \sum_{k=p_1+1}^{p_1+p_2} \eta + \delta_k] = (p_1 b\zeta + p_2\eta)^2\phi + p_1^2\zeta^2\sigma + p_2^2\eta^2\sigma \tag{77}$$

The Lagrangian for the problem is

$$L(\zeta, \eta, \Lambda) = (p_1 b\zeta + p_2\eta)^2\phi + p_1^2\zeta^2\sigma + p_2^2\eta^2\sigma - \Lambda(p_1\zeta^2 + p_2\eta^2 - 1) \tag{78}$$

The partial derivatives are

$$\frac{\partial L}{\partial \zeta} = 2\phi p_1 b(p_1 b\zeta + p_2\eta) + 2p_1\zeta\sigma^2 - 2\Lambda p_1\zeta = 0,$$
$$\frac{\partial L}{\partial \eta} = 2\phi p_2(p_1 b\zeta + p_2\eta) + 2p_2\eta\sigma^2 - 2\Lambda p_2\eta = 0, \tag{79}$$

Rearranging the terms leads to

$$\zeta = b\eta = b(p_1 b^2 + p_2)^{-1/2} \tag{80}$$

so the ratio of the weights is b, just as for the $\hat{\xi}$ in (63).

References

Anderson, T. W. (1963), 'Asymptotic theory for principal component analysis', *The Annals of Mathematical Statistics* **34**, 122–148.

Anderson, T. W. (2003), *An Introduction to Multivariate Statistical Analysis*, 3rd edn, John Wiley and Sons, New York.

Babakus, E., Ferguson, Jr., C. E. & Joereskog, K. G. (1987), 'The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions', *Journal of Marketing Research* **24**, 222–228.

Bai, J. (1993), 'Inferential theory for factor models of large dimensions', *Econometrica* **71**, 135–171.

Bartholomew, D. & Knott, M. (1999), *Latent Variable Models and Factor Analysis*, Kendall's Library of Statistics, 7, Arnold Publishers.

Bartolo, A. D. (2000), 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.

Bollen, K. (1989), *Structural Equations with Latent Variables*, Wiley and Sons, New York.

Bollen, K. A. & Barb, K. H. (1981), 'Pearson's R and coarsely categorized measures', *American Sociological Review* **46**, 232–239.

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

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

Bollen, K. A. & Long, J. S., eds (1993), *Testing Structural Equation Models*, SAGE Publications, Thousand Oaks, CA.

Brualdi, R. A. & Mellendorf, S. (1994), ‘Regions in the complex plane containing the eigenvalues of a matrix’, *The American Mathematical Monthly* **101**, 975–985.

Caudill, S. B., Zanella, F. C. & Mixon, F. G. (2000), ‘Is economic freedom one dimension? A factor analysis of some common measures of economic freedom’, *Journal of Economic Development* **25**, 17–40.

Choi, I. (2002), ‘Structural changes and seemingly unidentified structural equations’, *Econometric Theory* **18**, 744–775.

Corporation, S. (2003), *Stata Software, Release 8*, College Station, TX.

Davis, A. W. (1977), ‘Asymptotic theoty for principal component anakysis: Non-normal case’, *Australian Journal of Statistics* **19**, 206–212.

Demmel, J. W. (1997), *Applied Numerical Linear Algebra*, SIAM, Philadelphia.

DiStefano, C. (2002), ‘The impact of categorization with confirmatory factor analysis’, *Structural Equations Modeling* **9**, 327–346.

Dolan, C. V. (1994), ‘Factor analysis with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data’, *British Journal of Mathetmatical and Statistical Psychology* **47**, 309–326.

Drakos, K. (2002), ‘Common factor in eurocurrency rates: A dynamic analysis’, *Journal of Economic Integration* **17**, 164–184.

Filmer, D. & Pritchett, L. (1998), Estimating wealth effect without expenditure data — or tears: An application to educational enrollments in states of India, World Bank Policy Research Working Paper No. 1994, The World Bank, Washington, DC.

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

Flury, B. (1988), *Common Principal Components and Related Multivariate Methods*, John Wiley and Sons, New York.

Gwatkin, D. R., Rustein, S., Johnson, K., Suliman, E. A. & Wagstaff, A. (2003*a*), Socio-economic differences in health, nutrition, and population, Technical report, World Bank. Volume 1: Armenia – Kyrgyz Republic.

Gwatkin, D. R., Rustein, S., Johnson, K., Suliman, E. A. & Wagstaff, A. (2003*b*), Socio-economic differences in health, nutrition, and population, Technical report, World Bank. Volume 2: Madagascar – Zimbabwe.

Harris, D. (1997), 'Principal component analysis of cointegrated time series', *Econometric Theory* **13**, 529–557.

Hoeffding, W. (1948), 'A class of statistics with asymptotically normal distribution', *Annals of Mathematical Statistics* **19**, 293–325.

Horn, R. A. & Johnson, C. R. (1990), *Matrix Analysis*, Cambridge University Press, Cambridge, UK.

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

Huber, P. J. (2003), *Robust Statitsics*, John Wiley and Sons, New York.

Johnson, D. R. & Creech, J. C. (1983), 'Ordinal measures in mulitple indicator models: A simulation study of categorization error', *American Sociological Review* **48**, 398–407.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1997), *Discrete Multivariate Distributions*, John Wiley and Sons, New York.

Johnstone, I. M. (2001), 'On the distribution of the largest eigenvalue in principal component analysis', *Annals of Statistics* **29**, 295–327.

Jolliffe, I. T. (2002), *Principal Component Ananlysis*, 2nd edn, Springer, Heidelberg and New York.

Jöreskog, K. (2004*a*), 'Structural equation modeling with ordinal variables'.

Jöreskog, K. (2004*b*), *Structural Equation Modeling With Ordinal Variables using LISREL*. Notes on LISREL 8.52.

http://www.ssicentral.com/lisrel/ordinal.pdf.

Judd, K. L. (1998), *Numerical Methods in Economics*, MIT Press, Cambridge, MA.

Kaplan, D. (2000), *Structural Equation Modeling: Foundations and Extensions*, SAGE Publications, Thousand Oaks, CA.

Kendall, M. G. (1955), *Rank Correlation Methods*, 2nd edn, Charles Griffin & Co., London.

Kolenikov, S. (2001), ‘Review of Stata 7’, *Journal of Applied Econometrics* **16**, 637–646.

Krelle, W. (1997), How to deal with unobservable variables in economics, Discussion Paper B/414, Universitat Bonn.

Kullback, S. (1997), *Information Theory and Statistics*, Dover Publications, Mineola, NY.

Lebart, L., Morineau, A. & Warwick, K. M. (1984), *Multivariate Descriptive Statistical Analysis*, John Wiley and Sons, New York.

Maddala, G. S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*, Vol. 3 of *Econometric Society Monographs*, Cambridge University Press, Cambridge, UK.

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

Maydeu-Olivares, A. (2001), Testing categorized bivariate normality with two-stage polychoric correlation estimates, Technical report, University of Barcelona, Dept. of Psychology.

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

Parlett, B. (1980), *The Symmetric Eigenvalue Problem*, Prentice Hall, Englewood Cliffs, NJ.

Pearson, K. (1901*a*), ‘Mathematical contributions to the theory of evolution. vii. on the correlation of characters not qualitatively measurable’, *Philosophical Transactions of the Royal Society of London, Series A* **195**, 1–47.

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

Pearson, K. & Pearson, E. S. (1922), ‘On polychoric coefficients of correlation’, *Biometrika* **14**, 127–156.

Reichlin, L. (2002), Factor models in large cross-sections of time series, Discussion Paper DP3285, CEPR.

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

Skinner, C. J., Holmes, D. J. & Smith, T. M. F. (1986), 'The effect of sample design on principal component analysis', *Journal of the American Statistical Association* **81**, 789–798.

SSI (2004), *LISREL software, Release 8.52 for Windows*, Scientific Software International, Lincolnwood, IL.

Stock, J. H. & Watson, M. W. (2002), 'Forecasting using principal components from a large number of predictors', *Journal of the American Statistical Association* **97**, 1167–1179.

van der Vaart, A. W. (1998), *Asymptotic statistics*, John Wiley and Sons, New York.

Wansbeek, T. & Meijer, E. (2000), *Measurement Error and Latent Variables in Econometrics*, North-Holland, Amsterdam.

Webster, T. J. (2001), 'A principal component analysis of the U.S.News & World Report tier rankings of colleges and universities', *Economics of Education Review* **20**, 235–244.

Weisstein, E. W. (2004), 'Eigenvalue'. From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/Eigenvalue.html.

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, Cambridge, MA.