Econometric Reviews
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lecr20

Constructing Common Factors from Continuous and Categorical Data

Serena Ng [a]
[a] Department of Economics , Columbia University , New York , New York , USA
Accepted author version posted online: 03 Sep 2014.Published online: 17 Dec 2014.

To cite this article: Serena Ng (2015) Constructing Common Factors from Continuous and Categorical Data, Econometric Reviews, 34:6-10, 1141-1171, DOI: 10.1080/07474938.2014.956625

To link to this article: http://dx.doi.org/10.1080/07474938.2014.956625

Econometric Reviews, 34(6–10):1141–1171, 2015

ISSN: 0747-4938 print/1532-4168 online
DOI: 10.1080/07474938.2014.956625

Taylor & Francis
Taylor & Francis Group

Constructing Common Factors from Continuous and Categorical Data

Serena Ng
Department of Economics, Columbia University, New York, New York, USA

The method of principal components is widely used to estimate common factors in large panels of continuous data. This article first reviews alternative methods that obtain the common factors by solving a Procrustes problem. While these matrix decomposition methods do not specify the probabilistic structure of the data and hence do not permit statistical evaluations of the estimates, they can be extended to analyze categorical data. This involves the additional step of quantifying the ordinal and nominal variables. The article then reviews and explores the numerical properties of these methods. An interesting finding is that the factor space can be quite precisely estimated directly from categorical data without quantification. This may require using a larger number of estimated factors to compensate for the information loss in categorical variables. Separate treatment of categorical and continuous variables may not be necessary if structural interpretation of the factors is not required, such as in forecasting exercises.

Keywords Alternating least squares; Factor models; Ordinal data; Principal components.

JEL Classification C5; C6; C25; C35.

1. INTRODUCTION

The recent interest of economists in factor models is largely driven by the fact that common factors estimated from large panels of data often have predictive power for economic variables of interest. Theoretical and empirical work predominantly use principal components to estimate the common factors in continuous data. Little attention has been given to alternative estimators and to the treatment of categorical data even though many economic variables are of this nature. For example, households and firms are asked in surveys whether they expect economic conditions to improve or not. While such data could be useful for forecasting, they cannot be approximated by continuous distributions. This article first reviews dimension reduction methods that can handle mixed measurement data, meaning that the data can be continuous or categorical. I then

Address correspondence to Prof. Serena Ng, Department of Economics, Columbia University, 420 W. 118 St., New York, NY 10025, USA; E-mail: serena.ng@columbia.edu

investigate the consequence from the perspective of the factor space of using categorical variables to construct principal components, treating the data as if they were continuous.

Any study of socioeconomic status necessarily involves analyzing dichotomous data or data with a small number of categories. Such data requires special treatment as they contain important but imperfect information about the underlying latent variables. Racine and Li (2004) and Su and Ullah (2009) consider using a small set of mixed data in non-parametric regressions. Here, I consider the situation when the set of mixed predictors is large enough that dimenional reduction becomes necessary. As pointed out in Kolenikov and Angeles (2009) and further investigated below, the method that is currently used by economists is far from satisfactory.

The psychometric approach to the dimension reduction problem is to either explicitly model the latent continuous variables or quantify (impute) the continuous variables from the categorical data.[1] According to the Statistical Package for the Social Science (SPSS) software and as explained in Meulman and Heiser (2001), the following three types of categorical variables are relevant: (1) nominal variables which represent unordered categories (such as zip codes and SIC codes); (2) ordinal variables which represent ordered categories (such as satisfaction ratings of excellent/good/average/poor and Likert scale); and (3) numerical (count) variables which represent ordered categories (such as age in years and income class in dollars) with distances between categories that can be meaningfully interpreted. Nominal and ordinal data are said to be nonmetrical because the distance between two categories has no meaningful interpretation. The challenge for factor analysis of non-metrical data lies in the fact that normalization and monotonicity constraints need to be imposed to ensure consistency between the imputed variables and the observed discrete variables. Not surprisingly, going down this route necessarily takes us from linear to nonlinear methods of dimension reduction.

I begin with a review of factor analysis of continuous data from the viewpoint of solving a Procrustes problem. These methods are nonprobabilistic and do not permit formal inference to be made. But they form the basis of many dimension reduction problems which are interesting in their own right. The issues that arise in factor analysis of categorical data are then discussed. Special focus is given to methods that quantify the discrete data. Simulations are used to evaluate the precision of the factors estimated from continuous and mixed data. The so-called Filmer–Pritchett procedure is also evaluated. I assess the factors estimates from the perspective of diffusion index forecasting which requires extracting common information in a large number of categorical variables. Precise estimation of the factor space rather than structural interpretation of the factor estimates takes center-stage.[2] An interesting finding is that the principal components of the raw discrete data can estimate the factor space reasonably precisely, though this may

[1]A 1983 issue of *Journal of Econometrics* (de Leeuw and Wansbeek editors) was devoted to these methods.

[2]The focus is rather different from the structural factor analysis considered in Cunha and Heckman (2008) and Almund et al. (2011).

require overestimating the number of factors to compensate for the information loss in categorical data. Data quantification may not be necessary.

2. FACTOR ANALYSIS OF CONTINUOUS DATA

Factor analysis is a statistical framework used to analyze the behavior of observed variables using a small number of unobserved factors. Spearman (1904) appears to be the first to conjecture that a common unobserved trait (mental ability) may be responsible for the positive correlation in children's test scores on a variety of subjects. To analyze the contribution of the factors on the test scores and more generally on data that are continuous, the classical approach is to estimate the factor loadings by maximizing the Gaussian likelihood. Important contributions have subsequently been made by Anderson and Rubin (1956), Joreskog (1970), Lawley and Maxwell (1971), and Browne (1984), among others. Factor models are now used not just by psychologists, but by researchers in marketing, biology, and other fields.

Let X denote a $T \times N$ matrix of continuous data or data in ratio form. As a matter of notation, the i, j entry of X is denoted X_{ij}; $X_{i,:}$ is the ith row of X and $X_{:,j}$ is the jth column. For macroeconomic panels, N is the number of variables, and T is the number of time periods over which the variables are observed. A superscript zero is used to denote true values. The goal of factor analysis is to explain X using r common factors $F^0 = (F_1^0, \ldots, F_r^0)$ and N idiosyncratic errors e^0. In matrix form, the factor representation of the data is

$$X = F^0 \Lambda^{0\prime} + e^0,$$

where Λ^0 is a $N \times r$ matrix of factor loadings. The population covariance structure of X under the assumption that the factors have unit variance and are mutually uncorrelated is

$$\Sigma_X^0 = \Lambda^0 \Lambda^{0\prime} + \Omega^0.$$

In classical factor analysis, Ω^0 is assumed to be a diagonal matrix, and the data X are said to have a strict factor structure. Let k be the assumed number of factors which can be different from r. Anderson and Rubin (1956) assume that the data are multivariate normal. They use the normalization $\Sigma_{F^0} = I_r$ and suggest to estimate the factor loadings by maximizing the log likelihood:

$$\log\ L_0(\Lambda, \Omega; k) = \log|\Omega| + \text{trace}\ (X - F\Lambda')\Omega^{-1}(X - F\Lambda')'.$$

Lawley and Maxwell (1971) consider the equivalent problem of maximizing

$$\log L_1(\Lambda, \Omega; k) = \log|\Sigma_X| + \text{trace}\ (S_X \Sigma_X^{-1}) - \log|S_X| - N,$$

where S_X is the sample covariance of X. An advantage of maximum likelihood estimation is that the sampling distribution of $\widehat{\Lambda}$ is known and inference can be made. When T is large and N is fixed, the factor estimates $\hat{\Lambda}$ are $\sqrt{T}$ consistent and asymptotically normal. Unbiased estimates of F can be obtained as shown in Lawley and Maxwell (1971, Ch. 8) even though these are not usually the object of interest in psychology research.

When the normality assumption is not appropriate, one alternative is to consider covariance structure estimation (also known as structural equation modeling). Let $\theta = (\text{vec}\ (\Lambda), \text{vech}\ (\Omega))'$ be the parameters of the factor model and W be a weighting matrix. The weighted least squares estimator is

$$\hat{\theta}_{WLS} = \underset{\theta}{\text{argmin}}\ (\text{vech}\ (\Sigma_X(\theta) - \text{vech}\ (S_X))'W(\text{vech}\ (\Sigma_X(\theta) - \text{vech}\ (S_X)).$$

Under regularity conditions and assuming that N is fixed, the weighted least squares (WLS) estimator is also $\sqrt{T}$ consistent and asymptotically normal, as shown in Browne (1984) and others. As $\hat{\theta}_{WLS}$ is simply a method of moments estimator, it is less efficient than maximum likelihood estimator (MLE) but is robust to departures from normality. Assuming that the true number of factors is known, it has been documented that the asymptotic approximation of the WLS estimator is not accurate when the data exhibit excess kurtosis; the chi-square statistic for goodness of fit is oversized; the factor loadings are underestimated, and the standard error of the estimates tend to be downward biased.

Another alternative to MLE is iterative least squares estimation, the best known in this category being minimum residual (MINRES). Given a $N \times N$ sample correlation matrix R_X, the objective is to find a $N \times N$ matrix Ω and a $N \times k$ matrix Λ such that $\Lambda\Lambda'$ is as close to R_X as possible. Formally,

$$L_{MINRES}(\Lambda; k) = \left\| R_X - \Lambda\Lambda' - \Omega \right\|^2.$$

Concentrating out Ω and using the fact that the diagonal entries of R_X all equal one, the concentrated loss function is

$$L^c_{MINRES}(\Lambda; k) = \sum_{i \neq j} (R_{X,ij} - \Lambda_{i,:}\Lambda'_{j,:})^2,$$

where $\Lambda_{i,:}$ is the ith row of Λ, and $R_{X,ij}$ is the (i, j) entry of R_X. Harman and Jones (1966) suggest to start with an arbitrary Λ and iterate on each row of Λ holding other rows fixed. To update the rows, the objective function is separated into a part that depends on $\Lambda_{i,:}$ and a part c_i that does not. Let R_X^{-i} be the ith column of R_X with the i-element excluded, and let Λ^{-i} be the $(N-1) \times k$ matrix of Λ when the ith row is deleted. Define

$$\begin{aligned} L^c_{MRS,i}(\Lambda_{i,:}; k) &= (R_{X,1i} - \Lambda_{1,:}\Lambda'_{i,:})^2 + (R_{X,i-1,i} - \Lambda_{i-1,:}\Lambda'_{i,:})^2 \\ &\quad + (R_{X,i+1,i} - \Lambda_{i+1,:}\Lambda'_{i,:})^2 + \cdots + (R_{X,ki} - \Lambda_{k,:}\Lambda'_{i,:})^2 + c_i \\ &= \left\| R_X^{-i} - \Lambda^{-i}\Lambda'_{i,:} \right\|^2 + c_i. \end{aligned}$$

The solution to this minimization problem is standard; it is the least squares estimate

$$\widehat{\Lambda}'_{i,:} = (\Lambda^{-i'}\Lambda^{-i})^{-1}\Lambda^{-i'}R_X^{-i}.$$

Since all units other than i are held fixed when $\widehat{\Lambda}_{i,:}$ is constructed, the estimator is based on the principle of alternating least squares. Iterative updating of the ith row of Λ has been shown to decrease the loss function. However, a researcher would not be able to say whether a poor fit is due to the idiosyncratic errors or omitted factors. One shortcoming of MINRES is that $\widehat{\Lambda}_{i,:}\widehat{\Lambda}'_{i,:}$ can exceed one, which would then imply a negative idiosyncratic error variance. This so-called Heywood case can be circumvented. An example is the linear structural relations (LISREL) implementation of MINRES, detailed in Joreskog (2003).

Strict (or exact) factor models are to be contrasted with models more widely used in macroeconomics and finance. These models relax the assumption of strict factor models to allow some cross-section and serial correlation in e_{it}. Chamberlain and Rothschild (1983) referred to these as approximate factor models. Estimation is based on the method of asymptotic principal components (PCA) first proposed in Connor and Korajzcyk (1986). The estimator minimizes

$$L_{PCA}(\Lambda, F; k) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \Lambda_{i,:}F_t)^2 = \frac{1}{NT}\sum_{i=1}^{N}(X_{:,i} - F\Lambda'_i)'(X_{:,i} - F\Lambda'_i).$$

Because $\Lambda'_{i,:} = (F'F)^{-1}F'X_{:,i}$ for any given F, the concentrated objective function is

$$\begin{aligned} L^c_{PCA}(F; \Lambda; k) &= \frac{1}{NT}\sum_{i=1}^{N}(X_{:,i} - P_F X_{:,i})'(X_{:,i} - P_F X_{:,i}) \\ &= \frac{1}{NT}\sum_{i=1}^{N}X'_{:,i}X_{:,i} - \frac{1}{N}\sum_{i=1}^{N}X'_{:,i}P_F X_{:,i}. \end{aligned}$$

Since F and Λ are not separately identified, the normalization $F'F/T = I_k$ or $\Lambda'\Lambda/N = I_k$ is necessary. When $N > T$, minimizing $L^c_{PCA}(F; \Lambda)$ subject to the constraint that $F'F/T = I_k$ is the same as maximizing

$$\text{trace}\left(\frac{1}{N}\sum_{i=1}^{N}X'_{:,i}FF'X_{:,i}\right) = \text{trace}\left(F'\frac{1}{N}\sum_{i=1}^{N}X_{:,i}X'_{:,i}F\right) = \frac{1}{N}\text{trace}\,(F'XX'F).$$

The solution for F, denoted $\hat{F}$, is $\sqrt{T}$ times the eigenvectors corresponding to the k largest eigenvalues of the $T \times T$ matrix XX'. The solution for Λ, denoted $\hat{\Lambda}$, is $X'\hat{F}/T$. When $T > N$ and under the normalization that $\Lambda'\Lambda/N = I_k$, $\hat{\Lambda}$ is $\sqrt{N}$ times the eigenvectors corresponding to the k largest eigenvalues of the $N \times N$ matrix $X'X$ with $\hat{F} = X\hat{\Lambda}/N$. Note that unlike MLE and WLS which fix either T or N, the sample sizes in both dimensions are allowed to be large so that inferential theory can be developed

for both $\widehat{F}$ and $\widehat{\Lambda}$. Bai (2003) shows that if $\sqrt{N}/T \to 0$, then for each t, $\sqrt{N}(\hat{F}'_{t,:} - HF^{0\prime}_{t,:}) \xrightarrow{d} N(0, Avar(\hat{F}_t))$, and if $\sqrt{T}/N \to 0$, then for each i, $\sqrt{T}(\hat{\Lambda}'_{i,:} - H^{-1}\Lambda^{0\prime}_{i,:}) \xrightarrow{d} N(0, Avar(\hat{\Lambda}_i))$. As is well known, principal components describe but does not impose any structure on the data, while a factor model distinguishes between the common and idiosyncratic components. In a data rich environment, the distinction between factor models and principal components analysis is often left vague because the principal components consistently estimate the space spanned by the common factors.

A crucial assumption in the asymptotic theory for large dimensional factor models is that the largest eigenvalue of Ω^0 is bounded. A sufficient condition is that e^0_{it} is weakly correlated across i and t because under stationarity, the largest eigenvalue of Ω^0 is bounded by $\max_j \sum_{i=1}^N |Ee^0_{it}e^0_{jt}|$. Simulations show that $\widehat{F}_t$ precisely estimates the space spanned by F^0_t when Ω^0 is diagonal. This is a special case in which the eigenvalue bound is trivially satisfied. However, the construction of principal components as factor estimates do not take this constraint into account. Boivin and Ng (2006) report that estimated idiosyncratic errors associated with macroeconomic data tend to be quite pervasively cross-correlated. Onatski (2010) finds that asymptotic theory may not be a good guide to the finite sample properties of the factor estimates when the errors are strongly cross-sectionally correlated. However, few alternatives to the PCA are available. Doz et al. (2007) suggest a quasi-maximum likelihood approach that assumes Ω is diagonal even if the errors are serially and cross-sectionally correlated. A Kalman smoother is used to build up the likelihood which is then maximized using the expectation-maximization (EM) algorithm. They find that omitting the correlation in the errors has negligible effects on the estimated common factors when N is large. But their simulations only consider mildly correlated errors.

3. TWO ALTERNATING LEAST SQUARES ESTIMATORS (ALS)

As alternatives to the method of principal components, I consider two estimators (ALS1 and ALS2) that address the Haywood problem of negative idiosyncratic error variances. ALS2 additionally allows us to assess if the common and idiosyncratic components are poorly estimated. As distinct from all estimators considered in the previous section, the idiosyncratic errors are also objects of interest, putting them on equal footing with the common factors. Furthermore, the factors are estimated without writing down the probability structure of the model and as such, the statistical properties of the estimates are not known. It may perhaps be inappropriate to call these estimators. My interest in these methods arises because they can be extended to study latent variable models for mixed (discrete and continuous) data.

Whereas the standard approach to deriving an estimator is to take derivatives of the objective function, a derivative free alternative is to exploit information in the objective function evaluated at the upper or lower bound. For example, consider finding the minimum of $f(x) = x^2 - 6x + 11$. By completing the squares and writing $f(x) = (x - 3)^2 + 2$, it is clear that the lower bound for $f(x)$ is 2 and is achieved at $x = 3$. Knowing

the lower bound helps to determine the minimizer.[3] ten Berge (1993) and Kiers (2002) provide a formal treatment of using bounds to solve matrix optimizaiton problems.

Lemma 1. *Let Y be of order $p \times q$ with singular value decomposition $Y = PDQ'$. Let d_i be the ith diagonal value of D. Then (i) trace $(B'Y) \leq \sum_{i=1}^{q} d_i$ and (ii) the maximum of trace $(B'Y)$ subject to the constraint $B'B = I$ is achieved by setting $B = PQ'$.*

Kristof's upper bound for trace functions states that if G is a full rank orthonormal matrix and D is diagonal, then the upper bound of trace (GD) is $\sum_{i=1}^{n} d_i$. ten Berge (1993) generalizes the argument to sub-orthonormal matrices of rank $r \leq n$. Now trace $(B'Y) =$ trace $(B'PDQ') =$ trace $(Q'B'PD)$.[4] Part (i) follows by letting $G = Q'B'P$. Part (ii) says that the upper bound under the orthogonality constraint is attained by setting $G = I$ or equivalently, $B = PQ'$. This second result is actually the solution to the orthogonal "Procrustes problem" underlying many dimension reduction problems and is worthy of a closer look.[5]

Let A be a $n \times m$ matrix (of rank $r \leq m \leq n$), and let C be a $n \times k$ matrix with a specified structure. The orthogonal Procrustes problem looks for an orthogonal matrix B of dimension $m \times k$ that closely maps A to C. Formally, one chooses B to minimize $\|C - AB\|_F^2$ subject to the constraint $B'B = I_m$, where for a $p \times q$ matrix A, $\|A\|_F =$ trace $(AA')^{1/2} = \sqrt{\sum_{i=1}^{p} \sum_{j=1}^{q} A_{ij}^2}$.

By definition,

$$\|C - AB\|_F^2 = \text{trace } (C'C + B'A'AB - 2B'A'C).$$

When A has the same number of columns as C, B must be a $k \times k$ orthonormal matrix satisfying $B'B = BB' = I_k$. The first two terms do not depend on B since trace $(B'A'AB) =$ trace $(BB'A'A) =$ trace $(A'A)$. The minimization problem is thus the same as maximizing trace $(B'A'C)$ subject to the constraint that $B'B = I$. Letting PDQ' be the singular value decomposition of $A'C$ gives trace $(B'A'C) =$ trace $(B'PDQ') =$ trace $(Q'B'PD)$. Since $Q'B'P$ is orthogonal and D is a diagonal positive matrix, the trace is maximized if $Q'B'P$ is an identity matrix. The solution is to set $B = PQ'$. When $m > k$ and B is a long matrix with more rows than columns, the solution has no closed form and must be solved by iteration.

To see how Lemma 1 can be used to construct principal components, I follow the formulation in ten Berge (1993). Given a $T \times N$ matrix of standardized data X, the goal of principal components analysis is to find k linear combinations given by the $N \times k$ matrix

[3]It is important that the lower bound is attainable and does not depend on x. If the lower bound was 1 instead of 2, no meaning could be attached to $x = 3$ because the lower bound of 1 is not attainable.

[4]A matrix is sub-orthonormal if it can be made orthonormal by appending rows or columns.

[5]The orthogonal Procrustes problem was solved in Schonmenn (1966). See Gower and Dijksterhuis (2004) for a review for subsequent work.

B such that $F = XB$ optimally summarizes information in X. Let A' be the weights for estimating F from X. The loss function

$$L^c_{PCA}(B, A) = \left\| X - FA' \right\|^2$$

is to be minimized subject to $F'F/T = B'(X'X/T)B = B'R_XB = I_k$. Concentrating out $A' = (F'F)^{-1}FX$ and defining $Y = R_X^{1/2}B$ gives

$$\begin{aligned} L^c_{PCA}(B; A) &= \left\| X - XB(F'F)^{-1}F'X \right\|^2 = \left\| X - XBB'R_X \right\|^2 \\ &= \text{trace } (X'X) - 2T \text{ trace } (B'R_X^2B) + T \cdot \text{trace } (B'R_X^2B) \\ &= T \cdot k - T \cdot \text{trace } (B'R_X^2B) \\ &= T \cdot k - T \cdot \text{trace } (Y'R_XY). \end{aligned}$$

The constrained minimization problem now becomes maximizing trace $(B'R_XB)$ subject to $B'R_XB$, or equivalently, maximizing trace $(Y'R_XY)$ subject to $Y'Y = I_k$. From the eigen-decomposition that $R_X = PDP'$,

$$\text{trace } (Y'R_XY) = \text{trace } (Y'PDP'Y) = \text{trace } (P'YY'PD) = \text{trace } (GD)$$

with $G = P'YY'P$. The upper bound of trace $(GD) = \sum_{i=1}^k d_i$ is achieved if G is sub-orthonormal, i.e., $G = \left(\begin{smallmatrix} I_k & 0 \\ 0 & 0 \end{smallmatrix}\right)$. We thus let $Y = P_{:,1:r}$ and $B = P_{:,1:k}D_{1:k,1:k}$ (times any $k \times k$ orthonormal matrix). Canonical correlation analysis and many constrained least squares problems with no closed form solutions can also be solved using these monotonically convergent algorithms.

3.1. ALS1

I first consider the estimator proposed in De Leeuw (2004), Unkel and Trendafilov (2010) and Trendafilov and Unkel (2011). Let $e = u\Psi$ where $u'u = I_N$. Let k be the number of assumed (not necessarily equal r, the true) number of factors. Suppose first that $T \geq N$, and consider maximizing

$$\begin{aligned} &L_{ALS1}(F, \Lambda, u, \Psi; k) = \left\| X - F\Lambda' - u\Psi \right\|_F^2 \\ &\quad \text{subject to} \qquad (i)\ F'F = I_k, \qquad (ii)\ u'u = I_N, \qquad (iii)\ u'F = 0_{N\times k}, \\ &\qquad\qquad\qquad\quad (iv)\ \Psi \text{ is diagonal.} \end{aligned}$$

Notice that ALS minimizes the difference between the X and its fitted value $\widehat{X} = \widehat{F}\widehat{\Lambda}' + \hat{e}$. While the idiosyncratic and the common components are explicitly chosen, distributional assumptions on e are not made. Furthermore, the objective function takes

into account the off-diagonal entries of the fitted correlations. In contrast, the PCA loss function only considers the diagonal entries of $(X - F\Lambda)'(X - F\Lambda')$. Define

$$\underset{T\times(k+N)}{\underbrace{B}} = \begin{pmatrix} F & U \end{pmatrix}, \qquad \underset{N\times(k+N)}{\underbrace{A}} = \begin{pmatrix} \Lambda & \Psi \end{pmatrix}.$$

The ALS estimates are obtained by minimizing

$$L_{ALS1}(B, A; k) = \left\| X - BA' \right\|_F^2 \qquad \text{subject to } B'B = I_{N+k}.$$

But for given A, this is the same as maximizing trace $(B'XA)$ over B satisfying $B'B = I_{N+k}$. The problem is now in the setup of Lemma 1. When $N \leq T$, the estimates $(\tilde{F}, \tilde{\Lambda}, \tilde{U})$ can be obtained using the following three-step procedure:

1. Let $\widetilde{B} = PQ'$ where $\text{SVD}(XA) = PDQ'$.
2. From $\widetilde{B} = (\widetilde{B}_{:,1:k} | \widetilde{B}_{:,k+1:k+N})$, let $\widetilde{F} = \widetilde{B}_{:,1:k}$ and $\widetilde{U} = \widetilde{B}_{:,k+1:N}$. Update $\widetilde{\Lambda}$ as $X'\widetilde{F}$.
3. Let $\widetilde{\Psi} = \text{diag}(\widetilde{U}'X)$.

Steps (1) and (2) are based on part (ii) of Lemma 1. Step (3.) ensures that Ψ is diagonal and is motivated by the fact that

$$U'X = U'F\Lambda' + U'U\Psi.$$

Steps (1)–(3) are repeated until the objective function does not change. As $F\Lambda'$ is observationally equivalent to $FC^{-1}C\Lambda'$ for any orthogonal matrix C, Step (2) can be modified to make the top $k \times k$ submatrix of Λ lower triangular. This identification restriction does not affect the fact that rank $(\Lambda) = \min(\text{rank } (X), \text{rank } (F)) = k$.

When $N > T$, the rank of $U'U$ is at most T and the constraint $U'U = I_N$ cannot be satisfied. However, recall that $\Sigma_X = \Lambda\Lambda' + \Psi U'U\Psi$. Trendafilov and Unkel (2011) observe that the population covariance structure of the factor model can still be preserved if the constraint $\Psi U'U = \Psi$ holds, and in that case, $\Omega = \Psi^2$ is positive semidefinite by construction. Furthermore, the constraints $F'F = I_k$ and $U'F = 0_{N\times k}$ are equivalent to $FF' + UU' = I_T$ and rank $(F) = k$. Minimizing

$$L_{ALS1}(B, A; k) = \left\| X - BA' \right\|_F^2 \qquad \text{subject to } BB' = I_T$$

is the same as maximizing trace $(BA'X')$ which is again a Procrustes problem. The three step solution given above remains valid when $N > T$, but the singular value decomposition in Step (a) needs to be applied to $A'X'$.

Trendafilov and Unkel (2011) show that the objective function will decreases at each step and the algorithm will converge from any starting value. However, convergence of the objective function does not ensure convergence of the parameters. Furthermore, B is

not unique because it is given by the singular value decomposition of a rank deficient matrix. In particular, rank $(XA) \leq \min(\text{rank } (X), \text{rank } (A)) < \text{rank } (X) + k$.

3.2. ALS2

To motivate the second estimator, partition F^0 as $F^0 = \left(F^{01} \; F^{02}\right)$ where F^{01} has k columns and F^{02} has $r - k$ columns for some $k < r$. The data generating process can be written as

$$X = F^{01}\Lambda^{1\prime} + F^{02}\Lambda^{2\prime} + e^0 = F^{01\prime}\Lambda^1 + e^*.$$

If $k < r$ factors are assumed, the omitted factors F^{02} will be amalgamated with e^0 into e^*. Even though Ω^0 is diagonal in the population, the off-diagonal entries of the covariance matrix for e^* could be nonzero. Without explicitly imposing the restriction that X and e^0 are orthogonal, an estimator may confound e^0 with F_2^0. Socan (2003) reexamined an (unpublished) idea by H. Kiers to minimize

$$L_{ALS2}(F, \Lambda, e; k) = \left\|X - F\Lambda' - e\right\|_F^2$$
$$\text{subject to} \quad (i)\ e'F = 0 \quad (ii)\ e'e \text{ diagonal} \quad (iii)\ e'X \text{ diagonal}.$$

As with ALS1, estimation of e is explicitly taken into account. The first two constraints are standard; the third constraint ensures that the idiosyncratic errors are truly uncorrelated across units. This is because given orthogonality between e and F, any nonzero correlation between e_{it} and x_{jt} when $i \neq j$ can only arise if e_{it} is a function of the omitted factors.

The estimates $(\overline{F}, \overline{\Lambda}, \bar{e})$ are iteratively updated until the objective function does not change.

1. Given $\overline{\Lambda}$, let $\overline{J}$ be in the orthogonal null space of $\bar{e}$ so that $\overline{F} = \overline{JC}$ satisfies the constraints for $\overline{C} = J'X\Lambda(\Lambda'\Lambda)^{-1} = \min_C \|JC\Lambda' - X\|^2$.
2. For $i = 1, \ldots N$, update $\bar{e}_{:,i}$:

   a) Let $(X^{-i}, \bar{e}^{-i})$ be $(X, \bar{e})$ with the ith column set to zero;
   b) Let $H_i = (\, X^{-i} \;\; F \;\; \bar{e}^{-i}\,)$ with $\text{svd}(H_i) = P_{Hi}D_{Hi}Q_{Hi}'$;
   c) Let $\bar{e}_{:,i} = P_{Hi}^0 P_{Hi}^{0\prime} X_{:,i}$ where P_{Hi}^0 is the last column of P_{Hi}.

3. Update $\overline{\Lambda} = X'\overline{F}(\overline{F}'\overline{F})^{-1}$.

Steps (1) and (3) are intuitive as linear combinations of the basis in the orthonormal null space of e will be orthogonal to e. To find the optimal linear combinations, C is chosen to minimize $\|X - F\Lambda'\|^2 = \|X - JC\Lambda'\|^2$. This yields

$$\overline{C} = (\overline{J}'\overline{J})^{-1}\overline{J}'X\overline{\Lambda}(\overline{\Lambda}'\overline{\Lambda})^{-1} = \overline{J}'X\overline{\Lambda}(\overline{\Lambda}'\overline{\Lambda})^{-1}.$$

Step (2) is more involved. Recall that for each i, I need to find a $\bar{e}_{:,i}$ that it is uncorrelated with F and with $X_{:,j}$ and $\bar{e}_{:,j}$ for $j \neq i$. As rank (H_i) is less than its column dimension, P^0_{Hi} is orthogonal to all but the ith column of X and e as well as F. It remains to find the relation between $e_{:,i}$ and P^0_{Hi}. Minimizing $\|X_{,i} - P^0_{Hi}\beta_i\|$ yields $\bar{\beta}_i = P^{0\prime}_{Hi}X_{:,i}$. The update is thus $\bar{e}_{:,i} = P^0_{Hi}\bar{\beta}_i = P^0_{Hi}P^{0\prime}_{Hi}X_{:,i}$.

4. FACTOR ANALYSIS OF CATEGORICAL VARIABLES

Suppose the continuous data X are well represented by the factor model $X = F\Lambda' + e$, but we observe x which are discrete (nominal, ordinal, or numerical) transformations of X. Discrete data tend to be skewed and have excess kurtosis especially when the frequencies of observing a few categories are disproportionately high. Classical factor analysis is not appropriate because the normality assumption and Pearson correlation coefficients take as given that the data are continuous. As Kruskal and Shepard (1974) point out, a nonmetrical factor analysis needs to recover data that can be represented with the smallest number of factors and yet be consistent with the observed categorical data. Several approaches have been considered.

4.1. Probabilistic Approaches

The Bayesian approach is to parametrically specify the factor representation for X as well as the relation between X and x. This permits joint estimation of the factors and the loadings, as in Conti et al. (2012). Amongst the frequentist approaches, an "item response" analysis specifies the conditional distribution of the responses as a function of the latent factors assuming that the responses are independent conditional on the factors. A frequentist method that does not specify the mechanism that generates X is the "underlying variable approach". It simply assumes that each observed ordinal variable x_j is generated by a normally distributed latent variable X_j:

$$x_j = a_j \qquad \tau_j \leq X_j \leq \tau_{j-1}, \quad j = 1, \ldots C_j - 1,$$

where $\tau = (\tau_1, \ldots, \tau_{C_j-1})$ is a vector of $C_j - 1$ thresholds that define C_j categories. Both the item response and the underlying variable approach require calculations of tetrachoric or polychoric correlations.[6] These are correlations between two continuous variables

[6]Suppose that two continuous variables X_1 and X_2 are jointly normal with correlation coefficient ρ. The probability that $(X_1 > \tau_1, X_2 > \tau_2)$ is given by

$$p_{12}(\rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{\tau_1}^{\infty} \int_{\tau_2}^{\infty} \exp\Big(- \frac{y_1^2 + y_2^2 - 2\rho y_1 y_2}{2(1-\rho^2)} \Big) dy_1 dy_2.$$

The tetrachoric correlation proposed by Pearson (1900) is the ρ such that $p_{12}(\rho)$ equals the sample proportion $\hat{p}_{12}$. Polychoric correlations are then generalizations of tetrachoric correlations from two dichotomous indicators to multiple ordered class.

underlying the discrete indicators. Once they are computed, the likelihood or method of moments estimation can be used. Muthén (1984) proposes a three step estimator of the latent variable model as follows:

1. Standardize X to be mean zero, unit variance, and estimate the vector of thresholds τ from the univariate margins of the observed data;
2. Estimate the polychoric correlations from the bivariate margins of the observed variables given the thresholds estimates. Denote the matrix of polychoric correlations by $\widehat{\Sigma}_X$;
3. Estimate θ (the factor loadings and variances) by maximizing the log likelihood (expressed in terms of the observed frequencies) or by minimizing the function

$$\text{vech}\,(\widehat{\Sigma}_X - \Sigma_X(\theta))' W \text{vech}\,(\widehat{\Sigma}_X - \Sigma_X(\theta))$$

with respect to θ, where W is a positive definite weighting matrix.

Compared to the method WLS described earlier for continuous data, parametric assumptions and an additional step of estimating the polychoric correlations are necessary. Joreskog and Moustki (2001) compare the latent variable and item response methods and find that full information methods are not very practical even though they are theoretically appealing. These estimators are designed for ordinal data; they can be computationally demanding even when N is small. Fortunately, they are implemented in packaged software like LISREL, Joreskog and Sorbom (2006) and Joreskog (1994).

The properties of these estimators applied to ordered data are documented in Babakus et al. (1987), Muthen and Kaplan (1985), Dolan (1994), and Bollen (1989, Ch. 9), among others. The bias in the estimated loadings, while often less than 10% or so, depends on how far are the categorized data from normality. Furthermore, the covariance matrix of the estimated idiosyncratic errors often have nonzero off-diagonal entries even if the true idiosyncratic errors are mutually uncorrelated. The possibility that categorical data may lead to spurious factors has also been raised, McDonald (1985, Ch. 4).

4.2. FACTALS and Optimal Scaling

Methods have also been developed to estimate common factors in categorical data without specifying a probability model. Let X be a $N \times J$ matrix, and let x denote the observed but discretized values of X. Let Λ be a $N \times r$ matrix of factor loadings, $\Omega = D^2$ be a diagonal matrix of idiosyncratic error variances. If all J columns of X were continuous, the factor model implies the correlation structure $\Lambda\Lambda' + \Omega(D)$. MINRES then minimizes

$$L_{MINRES}(\Lambda, D; r) = \left\| R_X - \Lambda\Lambda' - \Omega(D) \right\|^2.$$

As pointed out earlier, the minimization problem can be solved columnwise because R_X is a correlation matrix. Now instead of X, a $N \times J$ matrix of data x is observed, some columns are of X continuous, and some are discrete. The challenge for nonmetrical factor analysis is that the discrete nature of the observables put constraints on the factor estimates. For example, to respect the fact that ordinal data are ordered, Kruskal and Shepard (1974) suggest to construct principal components subject to monotonicity constraints. Takane et al. (1979) argue that this does not fully exploit the factor structure. They suggest to replace R_X by R_Z, where Z is a $N \times J$ matrix of optimally scaled values of x.

Optimal scaling is an integral part of dimension reduction methods used to analyze nonmetrical variables. Let C_j be the number of categories in variable j, and let G_j be a $N \times C_j$ indicator matrix that is one if variable j is categorical with columns following the ordering of the categories. The adjacency matrix is defined as $G = [G_1, G_2, \ldots G_J]$. The quantified scores for variable j is $Z_j = G_j Y_j$ where Y_j is estimated subject to constraints of the measured data. For example, if $x_{:,j}$ is ordinal, the restriction that $Y_j(1) \geq Y_j(2) \geq \ldots Y_j(C_j)$ is required. The exercise is to iteratively estimate Y, Λ, and Ω by minimizing

$$L_{QFAC}(\Lambda, D, Z; k) = \left\| Z'Z - \Lambda\Lambda' - \Omega(D) \right\|^2$$

subject to

$$i)\ 1'Z_{:,j} = 0 \qquad (ii)\ Z'Z = I_J \qquad (iii) \text{ measurement level constraints}$$

Note that each column of Z is normalized to have unit sum of squares. This is a MINRES problem in which Z plays the role of X, and more importantly, Z is itself being estimated. In a sense, the analysis proceeds as though $Z_{:,j}$ has factor representation $Z_{:,j} = F\Lambda'_{j,:} + e_{:,j}$.

While the problem seems conceptually simple, it is not trivial computationally because this is a quadratic program with a quadratic constraint (in view of the normalization for $Z'Z$). Kiers et al. (1993) propose a monotonically convergent FACTALS (factor analysis by alternating least squares) algorithm for estimating k factors as follows[7]:

1. Let $\widehat{\Lambda} = U_{:,1:k} S_k^{1/2}$ where $Z'Z - \Omega$ has singular value decomposition USV, with $U_{:,1:k}$ being the first k columns of U;
2. Let $\widehat{\Omega} = \text{diag}(Z'Z - \widehat{\Lambda}\widehat{\Lambda}')$ (constrained to be non-negative);
3. Update $\widehat{R}_Z = \widehat{\Lambda}\widehat{\Lambda}' + \widehat{\Omega}$. For $j = 1, \ldots, J$:

   a) If $x_{:,j}$ is continuous, $Z_{:,j} = x_{:,j}$. Let $Z_{:,-j}$ be Z with the jth column removed, and $\widehat{R}_{Z,:,-j}$ be the jth column of $\widehat{R}_Z$ with the jth element removed;

[7]The initial procedure proposed by Takane et al. (1979) and refined by Nevels (1989) both have shortcomings. FACTALS fixes those bugs. Special thanks to H. Kiers for sharing the MATLAB code.

b) If $x_{j,:}$ is nominal, minimize $\|Z'_{:,-j}G_jy_j - \widehat{R}_{Z,:,-j}\|^2$ subject to the constraint that G_jy_j is centered and $y'_jG'_jG_jy_j = 1$. Given the solution y_j^0, update $Z_{:,j} = G_jy_j^0$;
c) If $x_{:,j}$ is ordinal, let $z = G_jy_j^0 + a^{-1}Z_{:,-j}\widehat{R}_{Z,:,-j} - a^{-1}Z_{:,-j}Z'_{:,-j}G_jy_j^0$, and minimize $\|z - G_jy_j\|^2$ subject to the constraints that (i) G_jy_j is centered, (ii) $y'_jG'_jG_jy_j = 1$, and (iii) the elements of y_j are weakly ordered. Given the solution y_j^0, update $Z_{:,j} = G_jy_j^0$;

4. Check if $\left\|Z'Z - \widehat{R}_Z\right\|^2$ converges. If not, return to step (1).

The thrust of FACTALS is to iteratively choose the scale values y_j to yield the quantified data Z_j and to update Λ and Ω. The first two steps perform columnwise update along the lines of MINRES. Step 3 imposes measurement level restrictions. Depending on the data type, it involves either solving an oblique Procrustes problem or performing a monotone regression. In a sense, FACTALS is a data augmentation method that treats X as missing values and imputes them as Z. These steps are further explained in the Appendix.

4.3. Principal Component Analysis

While economists rarely consider principal component analysis of qualitative data, the literature on this problem is in fact large. As surveyed in Michailidis and de Leeuw (1998), seemingly related research appears under a variety of names: homogeneity analysis, multiple correspondence analysis, PRINCALS systems, PRINCIPALS, discriminant analysis, to name a few.[8] Related approaches also include the principal factor analysis of Keller and Wansbeek (1983) and redudancy analysis (canonical correlation) of Israels (1984). As with continuous data, principal component analysis differs from factor analysis by going a step further to impose a structure.

Recall that given a $T \times N$ matrix of standardized continuous variables standardized to X, PCA computes $\widehat{\Lambda} = T^{-1}X'F$, where F is a $T \times r$ matrix of common components. PCA can be generalized to mixed data as follows. If variable j in the mixed data set is quantitative, let

$$S_j = \frac{1}{T}X_{:,j}X'_{:,j}$$

be the quantification matrix where $X_{:,j}$ contains the standardized values of the T observations on variable j. If variable j is qualitative, the quantification matrix is defined as

$$S_j = MG_jD_j^{-1}G_jM$$

[8]The method has been discovered and rediscovered under different names, including as quantification, multiple correspondence analysis, dual or optimal scaling, and homogeneity analysis. See Tenenhaus and Young (1985) for a synthesis of these procedures. However, none of these methods are familiar to economists.

where $M = I - 11'/T$ is the centering matrix, D_j is a diagonal matrix of frequencies of the categories in variable j, and G_j is the $T \times C_j$ indicator matrix for variable j. A principal components of mixed data then minimizes

$$\sum_{j=1}^{N} \text{trace } F'S_jF$$

over F subject to $F'F/T = I_r$. The solution is given by the first r eigenvectors of $\sum_j S_j$. The $T \times r$ matrix F then contains the standardized values of the components. If all variables are quantitative, the loadings are the eigenvectors of $X'X/T$, which is the PCA solution. If all variables are qualitative, the solution is the eigenvectors of $\sum_j MG_jD_j^{-1}M$. Sophisticated methods go one step further to impose level constraints (as in FACTALS) and may also allow for multiple quantification. See, for example, the PRINCIPALS routine in SAS and R, PRINCALS in R (HOMALS), and SPSS (CATPCA). I explore these methods in on-going empirical work but do consider them in simulations as these procedures are well studied in statistics and the psychometrics literature.

4.4. Experimental Alternatives

The difficulty in estimating and interpreting latent components from categorical data is that the population covariance of the categorical variables x (ie Σ_x) is not the same as the population covariance of the continuous variables X (ie. Σ_X). An estimate of the (i, j)th entry of Σ_x obtained by regressing $x_{:,i}$ on $x_{:,j}$. will be biased for the corresponding entry in Σ_X. Lancaster (1957) shows that, if a bivariate distribution is obtained by separate transformations of Y and Z that are bivariate normal, then the correlation of the transformed distribution cannot exceed the correlation coefficient of ρ_{YZ} in the bivariate normal distribution.[9]

In fact, discretization is a form of data transformation that is known to underestimate the linear relation between two variables, though the problem is alleviated as the number of categories increases. As mentioned earlier, many simulation studies have found that the r factor loadings estimated by MLE and WLS do not behave well when the data exhibit strongly non-Gaussian features. Data transformations can induce such features. But as seen above, estimating latent components from discrete data is quite not a trivial task.

A practical approach that has gained popularity in analysis of socioeconomic data is the so-called Filmer–Pritchett method used in Filmer and Pritchett (1998). Essentially, the method constructs principal components from the adjacency matrix, G. Kolenikov and Angeles (2009) assess the model's predicted rankings and find the method to be

[9]Olsson et al. (1982) show that ρ_{Yz} is downward biased for ρ_{YZ} if Y and Z are jointly normal. The greatest attenuation occurs when there are few categories and the data are opposite skewed. In the special case when consecutive integers are assigned to categories of Y, it can be shown that $\rho_{Yz} = \rho_{YZ} \cdot q$, where $q = \frac{1}{\sigma_\theta} \sum_{j=1}^{J-1} \phi(\alpha_j)$ and $\phi(\cdot)$ is the standard normal density and q is the categorization attenuation factor.

inefficient because it loses the ordinal information in the data. Furthermore, spurious negative correlation in the G matrix could underestimate the common variations in the data. However, they also find that constructing principal components from polychoric correlations of socio-economic data did not yield substantial improvements over the Filmer–Pritchett method.

I explore two alternatives that seem sensible when structural interpretation of the components in categorical variable is not necessary, and that N is large. The hope is that in such cases (as in economic forecasting), simpler procedures can be used to extract information in the categorical variables. The first idea is to construct the principal components from the quantified data. If principal components precisely estimates the space spanned by X, and Z are good quantifications of the discrete data x, then PCA applied to Z should estimate the space spanned by the common factors in X. I therefore obtain the quantified data $\widehat{Z}$ by FACTALS and then apply PCA to the covariance of $\widehat{Z}$ (i.e., $R_{\widehat{Z}}$) to obtain estimate $\widehat{F}$ and $\widehat{\Lambda}$. The approach is a natural generalization of the method of asymptotic principal components used when X was observed. However, the $\widehat{\Lambda}$ estimated by PCA applied to Z will generally be different from those that directly emerge from FACTALS because PCA does not impose diagonality of Ω. It will also be different from the ones that emerge from homogeneity analysis because level constraints are not imposed.

The second method is to ignore the fact that some variables are actually discrete and to extract principal components of Σ_x. As pointed out earlier, the eigenvectors of X will generally be different from those of x because $\Sigma_X \neq \Sigma_x$. In a sense, x is a contaminated copy of X. The hope is that the salient information in X will be retained in a sufficiently large number of eigenvectors of x, and that principal components will be able to extract this information. In other words, I compensate for the information loss created by data transformations with more latent components than would otherwise be used if X was observed.

5. MONTE CARLO SIMULATIONS

This section has two parts. The first subsection focuses on continuous data and assesses the precision of the factor estimates produced by the method of asymptotic principal components, ALS1, and ALS2. I also evaluate criterion for determining the number of factors. Subsection two turns to categorical and mixed data. All computations are based on MATLAB Release 2011a.

5.1. Continuous Data

Data with macroeconomic characteristics are generated from an approximate factor model. Specifically, two serially correlated factors are assumed with $\rho_{Fk} \sim U(0, 0.8)$ for

$k = 1, \ldots r$ and

$$X_{it} = \Lambda^0_{i,:} F^0_t + e_{it} \qquad \Lambda^0_{i,j} \sim N(0, \theta^2), \qquad \theta^0 = \{1, 0.5\}$$
$$F^0_{k,t} = \rho^0_{F,k} F^0_{k,t-1} + u_{kt}, \qquad u_{kt} \sim N(0,1)$$
$$e_{it} = \rho^0_e e_{it-1} + \varepsilon_{it}, \qquad \varepsilon'_{t,:} \sim N(0, I) C^0$$

with $\rho_e \sim U(0, 0.8)$. The degree of cross-section correlation in errors is deteremined by

$$C^0 = \text{toeplitz}([1, u_{1\times N_c}, 0_{1\times N - N_c - 1}]),$$

where $N_c = \{0, 0.1N\}$. As $\Omega^0 = C^0 C^{0\prime}$, the number of correlated series is much larger than $N/10$. The common component is strong when $\theta^0 = 1$ and weak if $\theta^0 = 0.5$, all else equal. Since Λ^0 is drawn randomly, F^0_1 is not necessarily more important than F^0_2. The relative importance of the common component in the population is given by

$$signal = \sum_{s=1}^{S} signal^s, \qquad signal^s = 1 - \frac{\sum_{i=1}^{N} \text{var}\ (e^s_i)}{\sum_{i=1}^{N} \text{var}\ (x^s_i)},$$

where s indexes the draw, and $S = 1{,}000$ is the number of replications. The properties of the estimates are judged by separately regressing $\widehat{F}_{1t}$ and $\widehat{F}_{2t}$ on a constant and the two dimensional F^0_t. The R^2 of the regressions indicate the coherence between $\widehat{F}_{kt}$ $(k = 1, 2)$ and the space spanned by the true factors.

Of the three estimators considered, the PCA is the easiest to compute as there is no iteration involved. While ALS1 converges in a few iterations, ALS2 is computationally the most demanding. It is the only estimator that sometimes (albeit rarely) fails to converge. Furthermore, the ALS2 estimator cannot be implemented when $N > T$ and I mark these estimates with a "-".

Table 1 reports results for the strong factor case with $\theta = 1$. The top panel has time dependent but cross-sectionally uncorrelated idiosyncratic errors since $N_c = 0$. With signal above 0.5, the three sets of estimated factors explain well over 0.95 of the variations in the true F_t. Assuming that e_{it} is cross-sectionally uncorrelated did not hurt the efficiency of PCA because the constraint is correct in this case.

The bottom panel of Table 1 allows the errors to be cross-sectionally correlated. As a consequence, the common component relative to the total variation in the data falls by as much as half. Notably, all factor estimates are less precise. One PCA factor tends to be more precisely estimated than the other. The discrepancy seems to increase as signal decreases. The two ALS estimators are much more even in this regard since $\widehat{F}_1$ and $\widehat{F}_2$ have similar predictive power of the factor space. Of the three estimators, ALS2 appears to be most unstable; it can be extremely good (such as when $(T, N) = (120, 80)$) or extremely bad (such as when T is increased to 240) holding N_c fixed at $= 8$. A factor that is precisely estimated by the PCA is not always precisely estimated by the ALS estimators

TABLE 1
R^2 from Regressions of $\widehat{F}_j$ on $F : \theta = 1$

T	N	N_c	*signal*	$\widehat{F}_1$	$\widehat{F}_2$	$\widehat{F}_1$	$\widehat{F}_2$	$\widehat{F}_2$	$\widehat{F}_2$
				PCA		*ALS I*		*ALS II*	
120	40	0	0.520	0.960	0.961	0.941	0.940	0.960	0.960
120	80	0	0.548	0.981	0.980	0.971	0.971	0.979	0.970
120	120	0	0.598	0.987	0.985	0.986	0.986	–	–
120	240	0	0.599	0.994	0.993	0.994	0.994	–	–
240	40	0	0.579	0.975	0.961	0.944	0.953	0.962	0.974
240	80	0	0.593	0.984	0.978	0.978	0.969	0.985	0.979
240	120	0	0.592	0.988	0.986	0.981	0.981	0.988	0.987
480	40	0	0.564	0.964	0.948	0.927	0.938	0.947	0.964
480	80	0	0.582	0.980	0.977	0.968	0.966	0.977	0.981
480	120	0	0.588	0.989	0.985	0.979	0.982	0.988	0.985
480	240	0	0.607	0.994	0.001	0.991	0.992	0.993	0.994
40	120	0	0.610	0.987	0.984	0.987	0.987	–	–
80	120	0	0.601	0.987	0.984	0.987	0.987	–	–
120	40	4	0.432	0.866	0.918	0.896	0.852	0.821	0.921
120	80	8	0.302	0.797	0.878	0.780	0.880	0.905	0.827
120	120	12	0.268	0.520	0.492	0.540	0.532	–	–
120	240	24	0.173	0.104	0.131	0.124	0.137	–	–
240	40	4	0.487	0.972	0.924	0.933	0.923	0.955	0.970
240	80	8	0.330	0.928	0.917	0.897	0.900	0.803	0.142
240	120	12	0.254	0.524	0.355	0.111	0.824	0.272	0.739
480	40	4	0.466	0.922	0.901	0.896	0.893	0.930	0.874
480	80	8	0.320	0.689	0.826	0.718	0.824	0.660	0.691
480	120	12	0.260	0.664	0.446	0.832	0.309	0.350	0.781
480	240	24	0.182	0.003	0.044	0.019	0.032	0.012	0.467
40	120	12	0.277	0.518	0.549	0.548	0.560	–	–
80	120	12	0.270	0.494	0.515	0.531	0.539	–	–

and vice versa. When $(T, N) = (240, 120)$, F_1^0 is poorly estimated by the ALS estimators (R^2 of 0.111 and 0.272) than by PCA (with R^2 of 0.524). However, the reverse is true of F_2^0, with R^2 of 0.824 and 0.739 for the ALS estimators, and only 0.355 for the PCA.

Table 2 considers weaker factor loadings with $\theta = 0.5$. When the errors are not cross-sectionally correlated, *signal* in the top panel of Table 2 is reduced somewhat relative to Table 1, but the correlation is not enough to strongly affect the precision of the factor estimates. For example, when $(T, N) = (120, 40)$, R^2 is 0.960 when $\theta = 1$ and is 0.865 when $\theta = 0.5$. When $(T, N) = (40, 120)$, R^2 goes from 0.987 to 0.94. When the idiosyncratic errors are also cross-sectionally correlated, the drop in signal is much larger. The R^2 values in the second panel of Table 2 are one-third to one-quarter of those in Table 1. Weak loadings combined with cross-correlated errors drastically reduce the precision of the factor estimates irrespective of the method used. When $(T, N) = (240, 120)$ which is not an unusual configuration of the data encountered in practice,

TABLE 2
R^2 from Regressions of $\widehat{F}_j$ on $F : \theta = 0.5$

T	N	N_c	signal	$\widehat{F}_1$	$\widehat{F}_2$	$\widehat{F}_1$	$\widehat{F}_2$	$\widehat{F}_1$	$\widehat{F}_2$
				PCA		*ALS I*		*ALS II*	
120	40	0	0.217	0.865	0.857	0.776	0.730	0.846	0.866
120	80	0	0.235	0.928	0.934	0.910	0.896	0.909	0.891
120	120	0	0.272	0.956	0.945	0.950	0.950	–	–
120	240	0	0.272	0.977	0.973	0.975	0.975	–	–
240	40	0	0.259	0.918	0.879	0.771	0.845	0.875	0.916
240	80	0	0.267	0.951	0.922	0.904	0.889	0.951	0.918
240	120	0	0.266	0.961	0.953	0.919	0.932	0.960	0.955
480	40	0	0.244	0.889	0.831	0.754	0.769	0.835	0.878
480	80	0	0.256	0.930	0.928	0.872	0.878	0.931	0.927
480	120	0	0.264	0.962	0.952	0.920	0.930	0.961	0.954
480	240	0	0.278	0.979	0.979	0.969	0.963	0.968	0.967
40	120	0	0.281	0.943	0.929	0.937	0.936	–	–
80	120	0	0.274	0.954	0.942	0.948	0.948	–	–
120	40	4	0.161	0.326	0.451	0.377	0.332	0.410	0.176
120	80	8	0.098	0.195	0.023	0.037	0.192	0.209	0.032
120	120	12	0.084	0.047	0.052	0.050	0.055	–	–
120	240	24	0.050	0.032	0.035	0.033	0.036	–	–
240	40	4	0.192	0.445	0.387	0.449	0.371	0.647	0.748
240	80	8	0.110	0.009	0.022	0.021	0.026	0.030	0.011
240	120	12	0.078	0.007	0.036	0.003	0.040	0.011	0.041
480	40	4	0.179	0.364	0.266	0.190	0.418	0.274	0.196
480	80	8	0.105	0.088	0.033	0.010	0.113	0.026	0.047
480	120	12	0.081	0.012	0.007	0.008	0.013	0.001	0.045
480	240	24	0.053	0.002	0.003	0.003	0.002	0.002	0.014
40	120	12	0.088	0.118	0.121	0.126	0.125	–	–
80	120	12	0.085	0.064	0.068	0.071	0.070	–	–

signal falls from 0.254 to 0.078 and the average R^2 drops from around 0.5 in Table 1 to less than 0.05 in Table 2! The difference is attributed to cross-correlated errors.

The results in Tables 1 and 2 are based on the assumption that r is known. Bai and Ng (2002) show that the number of factors can be consistently estimated by minimizing L_{PCA} subject to the constraint of parsimony. Specifically,

$$\hat{r}_{PCA} = \underset{k=k_{min},\ldots,k_{max}}{\operatorname{argmin}} \log L_{PCA}(k) + kg(N,T),$$

where $g(N,T) \to 0$ but $\min(N,T)g(N,T) \to \infty$. The ALS estimators are based on different objective functions. While their statistical properties are not known, Tables 1 and 2 find that the estimated ALS factors behave similarly to the PCA ones. I therefore let

$$\hat{r}_{ALS} = \underset{k=k_{min},\ldots,k_{max}}{\operatorname{argmin}} \log L_{ALS}(k)/nT + kg(N,T).$$

In the simulations, I use

$$g_2(N,T) = \frac{N+T}{NT}\log\min(N,T)$$

noting that $NT/(N+T) \approx \min(N,T)^{-1}$. This corresponds to IC_2 recommended in Bai and Ng (2008). For the ALS estimators I also consider a heavier penalty

$$g_A(N,T) = \frac{N+T}{NT}\log(N \cdot T).$$

The simulation design is similar to Tables 1 and 2. The criteria are evaluated for $k = 0,\ldots,6$. I only consider 500 replications because ALS2 is extremely time consuming to compute. The results are reported in Table 3. When $\theta = 1$ and the errors are cross-sectionally uncorrelated, the IC_2 almost always chooses the correct number of factors. The suitably penalized ALS objective functions also give the correct number of factors. The estimates of r are imprecise when there is cross-section dependence in e_{it}. The g_2 penalty often chooses the maximum number of factors whether PCA or ALS is used to estimate F, while the g_A tends to select too few factors. The results are to be expected; the penalties developed in Bai and Ng (2002) are predicated on a strong factor structure with weakly correlated idiosyncratic errors. When those assumptions are violated, the penalties are no longer appropriate.

We use AO to denote the test of Onatski (2010). His criterion exploits the square root shape of the edge of the eigenvalue distribution. The AO criterion of Onatski (2010) is supposed to better handle situations when there is substantial correlation in the errors. As seen from the third panel of Table 3, the AO criterion gives more precise estimates of r when the factor loadings are weak. However, it tends to select zero factors when many of the idiosyncratic errors are cross-sectionally correlated. Onatski (2010) argues that his criterion selects the number of factors that can be consistently estimated. It is not surprising that the AO criterion selects fewer factors when the factor component is weak. But taking the argument at face value would suggest that when signal is below 0.3, none of the two factors can be consistently estimated by the PCA or the ALS. This seems at odds with the fact that the estimated factors still have substantial correlation with the true factors.

Two conclusions can be drawn from these simulations. First, the objective function used to obtain $\widehat{F}$ seems to make little difference as the PCA and ALS estimates are similar. Second, not constraining Ω to be diagonal (as in the PCA) or unnecessarily imposing the constraint (as in the ALS) also does not have much effect on R^2. In this regard, the results echo those of Doz et al. (2007). If the strong factor assumptions hold true, there is little to choose between the estimators on the basis of the precise estimation of the factor space. Nonetheless, the PCA is computationally much less demanding.

Second, the precision of the factor estimates are strongly influenced by weak factor loadings. While signal is not observed and it is not known if the factors are strong or weak

TABLE 3
Estimates of $r = 2$ Using Continuous Data

T	N	N_c	signal	g_2	g_2	g_A	g_2	g_A	AO
		$\theta = 1$		PCA		*ALS1*		*ALS*	*PCA*
120	40	0	0.497	2.002	2.006	2.000	1.998	1.010	2.040
120	80	0	0.495	2.000	2.000	2.000	2.142	1.634	2.072
120	240	0	0.495	2.000	2.000	2.000	3.378	1.746	2.006
240	40	0	0.490	2.000	2.026	2.000	1.996	1.022	2.020
240	80	0	0.489	2.000	2.000	2.000	2.006	1.932	2.030
240	120	0	0.491	2.000	2.000	2.000	2.002	1.998	2.022
480	40	0	0.486	2.000	2.026	2.000	2.004	1.010	2.010
480	80	0	0.488	2.000	2.000	2.000	2.012	1.966	2.026
480	120	0	0.487	2.000	2.000	2.000	2.002	1.998	2.020
480	240	0	0.488	2.000	2.000	2.000	2.004	2.004	2.006
120	40	4	0.423	6.000	5.732	1.448	1.912	1.000	1.838
120	80	8	0.300	6.000	5.998	1.008	4.782	1.002	0.276
120	240	24	0.165	6.000	6.000	1.002	5.242	1.092	0.010
240	40	4	0.418	6.000	6.000	1.622	1.994	1.000	1.908
240	80	8	0.300	6.000	6.000	1.112	5.860	1.002	0.184
240	120	12	0.239	6.000	6.000	4.600	5.966	1.000	0.004
480	40	4	0.415	6.000	6.000	1.712	2.022	1.000	1.946
480	80	8	0.300	6.000	6.000	1.970	5.920	1.002	0.132
480	120	12	0.236	6.000	6.000	6.000	5.986	1.000	0.000
480	240	24	0.162	6.000	6.000	6.000	5.992	5.922	0.000
			$\theta = 0.5$						
120	40	0	0.086	0.004	1.010	1.000	1.000	1.000	0.488
120	80	0	0.085	0.000	1.000	1.000	1.300	1.000	0.918
120	240	0	0.086	0.658	1.000	1.000	1.124	1.000	1.960
240	40	0	0.084	0.006	1.384	1.000	1.000	1.000	1.164
240	80	0	0.084	0.096	1.614	1.000	1.000	1.000	1.930
240	120	0	0.084	0.606	1.682	1.000	1.002	1.000	2.016
480	40	0	0.083	0.008	1.940	1.006	1.000	1.000	1.888
480	80	0	0.083	0.396	2.000	1.064	1.000	1.000	2.014
480	120	0	0.083	1.452	2.000	1.178	1.000	1.000	2.020
480	240	0	0.083	2.000	2.000	1.440	1.008	1.000	2.008
			$\theta = 0.25$						
120	40	0	0.244	1.938	2.004	1.280	1.010	1.000	2.034
120	80	0	0.242	2.000	2.000	1.324	1.618	1.024	2.072
120	240	0	0.243	2.000	2.000	1.092	1.406	1.022	2.004
240	40	0	0.239	1.982	2.004	1.924	1.064	1.000	2.066
240	80	0	0.238	2.000	2.000	2.000	1.662	1.002	2.022
240	120	0	0.239	2.000	2.000	2.000	1.958	1.052	2.012
480	40	0	0.236	1.996	2.002	2.000	1.094	1.000	2.038
480	80	0	0.237	2.000	2.000	2.000	1.874	1.000	2.042
480	120	0	0.237	2.000	2.000	2.000	1.962	1.078	2.044
480	240	0	0.237	2.000	2.000	2.000	2.032	1.576	2.008

in practice, two indicators can be useful. The first is R^2, which should increase with signal. A low R^2 in spite of using many factors would be a cause for concern. The second is the discrepancy between the number of factors selected by IC_2 and AO. The two estimates should not be far apart when the factor structure is strong. When the $\hat{r}$s are very different, the strong factor assumptions may be questionable.

5.2. Mixed Data

Two designs of categorical data are considered. In the first case, the ordinal data x consists of answers by N respondents (such as professional forecasters) to the same question (such as whether they expect inflation to go up, down, or stay the same) over time T periods. In the second case, x consists of J responses (such as on income and health status) for each of the N units (such as households). In these experiments, PCA is used to estimate $\hat{r}$ factors in (i) the continuous data X as if it were observed, (ii) the categorical data x, (iii) the adjacency matrix G, and (where appropriate) (iv) the quantified data Z. The number of factors is determined by the criterion in Bai and Ng (2002) with penalty g_2 or the AO test of Onatski (2010). I begin with the first case when all data are ordinal.

a) PCA of X, x : and G. Data on one variable X are generated for N units and T time periods. The $J-1$ thresholds are spaced so that the probability of being in each interval is the same, and are generated using NORMINV (0.05:J–1:0.95). The data matrix is first standardized columnwise and then categorized into x which has J groups.

The results are given in Table 4. Columns 3 and 4 show that if X was observed, the number of factors would be precisely estimated. As seen from columns 5 and 6, $\hat{r}$ remains fairly precisely estimated when the categorical data x are used instead of X. However, the estimated number of factors in the adjacency matrix G is less stable. There are too few factors when the sample size is small but too many factors when N and T are large.[10]

Turning to an assessment of the estimated factor space, R^2_X indicates the average R^2 when $\hat{r}$ principal components are estimated from X where the $\hat{r}$ is determined by penalty g_2. The interpretation is similar for R^2_x and R^2_G. Evidently, $\widehat{F}$ precisely estimates F when X was available for analysis. The R^2s are slightly lower if the factors are estimated from x but the difference is quite small. The average R^2 remains well over 0.95. However, the principal components of the adjacency matrix G are less informative about the true factors. When the sample size is small and $\hat{r}$ underestimates r, R^2_G can be much lower than R^2_x. For example, when $r=2$, R^2_x is 0.916 when $(T,N)=(100,20)$, but R^2_G is only 0.368. The situation improves when the sample size increases as estimating more factors in G compensates for the information loss in the indicator variables. However, even with large $\hat{r}$, the principal components of G remain less informative about F^0 than the principal components of x. When $(T,N)=(200,100)$, R^2_x is 0.956 while R^2_G is 0.849, even though on average, $\hat{r}=3>r=2$ factors are found in G.

[10]In an earlier version of the article when x and G were not demeaned, PCA estimated one more factor in both x and G.

TABLE 4
Estimates of r from Ordinal Data

		Number of Factors						R^2		
T	N	IC_X	AO_X	IC_x	AO_x	IC_G	AO_G	R^2_X	R^2_x	R^2_G
					$r = 1$					
50	20	1.050	1.014	1.000	1.014	0.854	0.914	0.959	0.927	0.557
100	20	1.052	1.006	1.000	1.026	1.140	1.156	0.959	0.929	0.676
200	20	1.060	1.008	1.000	1.024	1.342	1.310	0.959	0.928	0.687
50	50	1.000	1.010	1.000	1.026	1.136	1.606	0.984	0.953	0.777
100	50	1.000	1.010	1.000	1.018	1.502	1.962	0.984	0.955	0.791
200	50	1.000	1.006	1.000	1.010	1.890	2.144	0.984	0.954	0.796
50	100	1.000	1.004	1.000	1.018	1.244	2.016	0.992	0.963	0.826
100	100	1.000	1.010	1.000	1.006	1.872	2.368	0.992	0.962	0.830
200	100	1.000	1.016	1.000	1.058	2.006	2.864	0.992	0.962	0.831
					$r = 2$					
50	20	2.714	2.024	2.000	2.028	0.324	0.524	0.958	0.919	0.119
100	20	3.302	2.032	2.000	2.020	1.126	1.004	0.960	0.916	0.368
200	20	3.424	2.014	2.000	2.010	1.916	1.458	0.959	0.917	0.577
50	50	2.000	2.034	2.000	2.016	1.236	1.502	0.984	0.949	0.488
100	50	2.000	2.016	2.000	2.024	2.090	2.344	0.984	0.947	0.750
200	50	2.000	2.010	2.000	2.016	2.784	2.902	0.984	0.947	0.797
50	100	2.000	2.014	2.000	2.022	1.742	2.138	0.992	0.958	0.708
100	100	2.000	2.004	2.000	2.062	2.710	2.630	0.992	0.957	0.841
200	100	2.000	2.018	2.000	2.184	3.092	2.842	0.992	0.956	0.849
					$r = 3$					
50	20	4.212	3.012	2.998	3.020	0.082	0.398	0.960	0.908	0.022
100	20	4.450	3.002	3.000	3.018	0.330	0.518	0.960	0.906	0.085
200	20	4.478	3.002	3.000	3.004	1.040	0.852	0.960	0.906	0.249
50	50	3.006	3.032	3.000	3.012	0.336	0.716	0.984	0.944	0.103
100	50	3.028	3.012	3.000	3.030	1.358	1.704	0.984	0.943	0.382
200	50	3.138	3.014	3.000	3.036	2.756	2.776	0.984	0.942	0.686
50	100	3.000	3.006	3.000	3.032	0.760	1.290	0.992	0.955	0.243
100	100	3.000	3.014	3.000	3.018	2.750	2.106	0.992	0.954	0.751
200	100	3.000	3.008	3.000	3.054	3.930	2.756	0.992	0.953	0.868

For $y = X$ (latent continuous data), x (categorical data), and G (adjacency matrix of indicators), IC_y denotes the number of factors selected by the Bai and Ng (2002) criterion with penalty $g_2 = \frac{N+T}{NT}\log\min(N,T)$ when the principal components are constructed from data y. AO_y denotes factors determined using the criterion of Onatski (2010). The columns R^2_y denote the average R^2 when each of the $\hat{r}$ factors estimated from y are regressed on the true factors.

b) PCA of X, x, G, and Z. Data for J variables for each of the N units are generated as

$$X_{ij} = \Lambda^0_{i,:} F^0_t + e_{ij},$$

where $e_{ij} \sim N(0, \sigma^2)$, $\Lambda^0_{i,:}$ is a $1 \times r$ vector of standard normal variates, $F^0_t \sim N(0, I_r)$. The factor loadings are $N(0, \theta^2)$. The J continuous variables are categorized using unevenly

TABLE 5
Ordinal Data, $r = 2$

N	J	n_G	$\hat{r}_X$	$\hat{r}_G$	$\hat{r}_Z$	$\hat{r}_x$	R^2_X	R^2_G	R^2_Z	R^2_x
				$\theta = 1$						
100	5	18	4.000	3.513	4.000	4.000	0.742	0.475	0.609	0.630
100	10	33	3.828	2.648	3.013	3.993	0.798	0.497	0.650	0.686
100	15	50	3.332	2.346	2.646	3.824	0.832	0.544	0.690	0.723
200	5	18	4.000	3.777	4.000	4.000	0.734	0.481	0.599	0.616
200	10	33	3.806	2.971	3.051	3.994	0.793	0.527	0.652	0.675
200	15	50	3.314	2.677	2.688	3.847	0.830	0.590	0.701	0.715
100	20	60	2.025	1.945	1.898	3.569	0.928	0.744	0.783	0.811
100	30	80	2.013	2.001	1.848	3.693	0.941	0.778	0.770	0.821
200	20	60	2.015	2.209	1.966	3.695	0.928	0.787	0.817	0.811
200	30	80	2.007	2.305	1.958	3.814	0.941	0.811	0.819	0.818
					$\theta = 0.5$					
100	5	18	4.000	2.212	4.000	4.000	0.487	0.219	0.384	0.400
100	10	33	2.644	1.289	2.237	4.000	0.462	0.164	0.269	0.446
100	15	50	1.433	0.331	0.436	2.096	0.562	0.115	0.155	0.458
200	5	18	4.000	2.594	4.000	4.000	0.469	0.234	0.369	0.377
200	10	33	2.663	1.600	2.228	4.000	0.472	0.205	0.270	0.430
200	15	50	1.556	0.735	0.431	2.159	0.602	0.240	0.155	0.470
100	20	60	1.641	0.423	0.453	2.710	0.669	0.155	0.165	0.529
100	30	80	1.769	0.488	0.412	3.159	0.744	0.184	0.153	0.562
200	20	60	1.788	0.973	0.454	3.051	0.723	0.335	0.168	0.554
200	30	80	1.881	1.133	0.442	3.485	0.786	0.399	0.168	0.578

For $y = X, x, G, Z$ where Z denotes quantified data, r^2_y denotes the number of factors estimated by the IC criterion of Bai and Ng (2002) with penalty g_2. R^2_y is the average R^2 when each of the factors estimated from y is regerssed on all the true factors.

spaced thresholds as summarized in Table 5a. The total number of categories (also the dimension of G) is denoted n_G.

TABLE 5a
Ordinal Variables

Ordinal Series	*Thresholds*	*Series*	*Thresholds*
1	−1.5, −0.75, 0.75, 1.5	2	−0.5 0.5
3	0	4	−0.5 0 0.5
5	−0.4 0.4 1	6	−1 0 1
7	0.4 1 0.6	8	0.3
9	1	10	0.6 1.2
11	0.2 0.6	12	−0.2 0.6
13	−0.7 0 0.7	14	−1.2 0.2 1.5
15	0.75 1.2		

The results for the case of strong loadings ($\theta = 1$) are in the top panel of Table 5. As seen from columns 5 to 8, the number of factors is overestimated whenever J is small even if X was observed; the average R^2 is 0.75 when J is 5 but increases to 0.941 when $J = 30$. This just shows that principal components can precisely estimate the factor space only when J is reasonably large. There is undoubtedly a loss of precision when X is not observed. The principal components of G yield low R^2 for small values of J. Apparently, G contains less information about F than Z. While the number of factors found in the discrete data x exceeds r, the $\hat{r}$ principal components of x give R^2s close to those based on Z. In other words, with enough estimated factors, the factor space can be as precisely estimated from discrete data than as from the quantified data.

Table 5a are results for the case of weak loadings with $\theta = 0.5$. Now R^2_Z and R^2_G are much reduced as the corresponding number of factors found in Z and G tends to be below the true value of two. This suggests that when the factor structure is already weak, discretization further weakens the information about the common factors in x. It is thus difficult to recover F from quantified data or from transformations of x. In such a case, the principal components of the raw categorical data x give the most precise factor estimates, and they are easiest to construct.

The final set of simulations consider mixed continuous, nominal and ordinal data. The number of continuous variables is always 20, the number of nominal variables $J_{nominal}$ is

Nominal Variables			
Series	*Thresholds*	*Series*	*Thresholds*
1	−1.5, −0.75, 0.75, 1.5	2	−0.2 0.5 1
3	0.8	4	−0.5 0.5
5	0.4 1	6	−1 0 1
7	0.5 0.5	8	−0.3 0.3
9	−1 1	10	−1.2 1.2

Ordinal Variables			
Series	*Thresholds*	*Series*	*Thresholds*
1	0.2 0.6	2	−0.2 −0.6
3	−.7 0 0.7	4	−1.2 0.2 1.5
5	−.75 1.2	6	−1.3
7	−1	8	−0.7
9	0	10	0.3
11	−1.5, −0.75, 0.75, 1.5	12	−0.5 0.5
13	0	14	−0.5 0 0.5
15	−0.4 0.4 1	16	−1 0 1
17	−0.7 0.4 1.6	18	−0.3
19	1	20	−0.6 1.2

TABLE 6
Mixed data, $r = 2$

N	$J_{nominal}$	$J_{ordinal}$	n_G	$\hat{r}_X$	$\hat{r}_G$	$\hat{r}_Z$	$\hat{r}_x$	R^2_X	R^2_G	R^2_Z	R^2_x
					$\theta = 1$						
100	5	5	31	2.000	4.000	2.000	2.003	0.953	0.491	0.940	0.941
100	10	5	42	2.000	4.000	2.000	2.002	0.957	0.497	0.942	0.943
100	5	10	41	2.000	3.997	2.000	2.003	0.958	0.585	0.942	0.943
200	5	5	31	2.000	4.000	2.000	2.003	0.953	0.470	0.942	0.941
200	10	5	42	2.000	4.000	2.000	2.002	0.956	0.478	0.944	0.943
200	5	10	41	2.000	4.000	2.000	2.004	0.958	0.542	0.944	0.943
200	5	20	74	2.000	3.615	2.000	2.004	0.961	0.608	0.946	0.945
200	10	20	85	2.000	3.467	2.000	2.003	0.963	0.646	0.948	0.947
400	5	5	31	2.000	4.000	2.000	2.010	0.953	0.468	0.943	0.942
400	10	5	42	2.000	4.000	2.000	2.007	0.957	0.471	0.945	0.944
400	5	10	41	2.000	4.000	2.000	2.014	0.958	0.528	0.945	0.944
400	5	20	74	2.000	3.716	2.000	2.012	0.961	0.595	0.947	0.946
400	10	20	85	2.000	3.669	2.000	2.010	0.963	0.637	0.949	0.947

either 5 or 10, and the number of ordinal variables $J_{ordinal}$ is 5, 10, or 20.[11] The results are given in Table 6. The number of factors is correctly estimated from the continuous data X, the quantified data Z, or the discrete data x but are overestimated from the dummy variable matrix G. The factor space is precisely estimated using X, Z or x but not G. The results are similar to those for purely ordinal data.

6. CONCLUSION

This article reviews and explores various matrix decomposition based methods for estimating the common factors in mixed data. A few conclusions can be drawn. First, if all data are continuous and T and N are large, ALS estimators have no significant advantage over PCA which is much simpler to construct. Second, with mixed data, the principal components of G give the least precise estimates of the factor space. Third, FACTALS provides a way to quantify the data consistent with a factor structure. But while the factor space can be precisely obtained from Z when the factor structure is strong, they are no more precise than analyzing x directly, and it is not robust when the factor structure in X is weak. Fourth, the observed categorical data x can be used to estimate the factor space quite precisely though additional factors may be necessary to compensate for the information that is lost from data discretization. It should be emphasized that different conclusions might emerge if criterion other than the factor space is used. Furthermore, the

[11]The FACTALS has convergence problems when N is 100 and the dimension of G is large.

Monte Carlo exercise is quite limited in scope. Nonetheless, the findings are encouraging and merit further investigation.

I consider FACTALS because it constructs quantified data Z consistent with a factor structure which allows me to consider a large dimensional factor analysis treating Z as data. However, as pointed out earlier, the sample correlations of transformed data are not the same as the raw data. Nothing ensures that Z will necessarily recover X from x. Thus, an important caveat of nonmetrical factor analysis is that Z may not be unique even when its second moments are consistent with the factor structure implied by the data X. Buja (1990) warns that the transformed data may identify spurious structures because singular values are sensitive to small changes in the data. A better understanding of these issues is also necessary.

7. APPENDIX

This appendix elaborates on how nominal and ordinal data are optimally scaled. It will be seen that the steps rely on matrix arguments presented in Section 3.

Step 3b: Nominal Variables. Observe first that $Z'_{:,-j}G_j y_j = Z'_{:,-j}Z_{:,j}$ defines the vector of "sample" cross correlations, while $\widehat{R}_{Z,-j}$ is the model implied analog. The j-entry is omitted since it is one by construction in both the sample and the model. We put sample in quotes because Z are quantified variables and not data.

Let $M = I_T - 11'/T$ be the idempotent matrix that demeans the data. Constraining Z to be mean zero columnwise is equivalent to imposing the condition $MG_j y_j = G_j y_j$ for every j with $Z'_{:,-j}G_j y_j = Z'_{:,-j}MG_j y_j$. Let B be an orthonormal bases for $MG_j y_j$. Then $MG_j y_j = B\tau$ for some τ. Instead of minimizing $\|Z'_{:,-j}G_j y_j - \widehat{R}_{Z,:,-j}\|^2$ over y_j subject to the constraint $y'_j G'_j G_j y_j = y'_j G'_j MG_j y_j$, the problem is now to minimize

$$\left\|Z'_{-j,:}B\tau - \widehat{R}_{Z,:,-j}\right\|^2$$

over τ subject to the constraint that $\tau' B' B\tau = \tau'\tau = 1$. This is an oblique Mosier's Procrustes problem whose solution, denoted τ_0, is given in Cliff (1966) and ten Berge and Nevels (1997). Given τ_0, y_j can be solved from $MG_j y_j = B\tau_0$. By least squares argument,

$$y_j^0 = (G'_j G_j)^{-1} G'_j B\tau_0.$$

It remains to explain the oblique Procrustes problem. Recall that the orthogonal Procrustes problem looks for a $m \times k$ transformation matrix B to minimize $\|\Lambda - AB\|^2$ subject to $B'B = I$. The oblique Procrustes problem imposes the constraint $\text{diag}(B'B) = I$. Since the constraint is now a diagonal matrix, each column of B can be solved separately. Let β be a vector from B, and λ be a column of Λ. The objective is now to find β

to minimize $(\lambda - A\beta)'(\lambda - A\beta)$. For U such that $UU' = U'U = I$ and $A'A = UCU'$, the problem can be rewritten as

$$(\lambda - A\beta)'(\lambda - A\beta) = (\lambda - AU\beta)'(\lambda - AU\beta).$$

By letting $w = U'\beta$ and $x = U'A\lambda$, the problem is equivalent to finding a vector w to minimize

$$\lambda'\lambda - 2x'w + w'Cw \qquad \text{subject to } w'w = 1.$$

The objective function and the constraint are both in quadratic form. The (nonlinear) solution depends on q, the multiplicity of the smallest latent root of $A'A$. In the simulations, I use the algorithm of ten Berge and Nevels (1997). In the present problem, $\lambda = \widehat{R}_{Z,:,-j}$ and $A = Z'_{:,-j}B$.

Step 3c: Ordinal Variables. When $X_{:,j}$ is ordinal, we minimize a function that majorizes $f(y_j)$ and whose solution is easier to find. As shown in Kiers (1990), this is accomplished by viewing f as a function of $q = G_j y_j$. The function of interest (for given q^0)

$$f(q) = \left\| Z'_{:,-j}q - \widehat{R}_{Z,:,-j} \right\| = \widehat{R}'_{Z,:,-j}\widehat{R}_{Z,:,-j} - 2\widehat{R}'_{Z,:,-j}Z'_{:,-j}q + \text{trace } (Z_{:,-j}Z'_{:,-j}qq'),$$

is majorized by

$$g(q) = c_1 + a\left(\left\| q^0 - (2a)^{-1}(-2Z_{:,-j}\widehat{R}_{Z,:,-j} + 2Z_{:,-j}Z'_{:,-j}q^0) - q \right\|^2 + c_2 \right)$$

where c_1 and c_2 are constants for q, and a is the first eigenvalue of $Z_{:,-j}Z'_{:,-j}$. Reexpressing q_j in terms of $G_j y_j$, we can maximize

$$\begin{aligned} h(y_j) &= \left\| G_j y_j^0 - (2a)^{-1}(-2Z_{:,-j}\widehat{R}_{Z,:,-j} + 2Z_{:,-j}Z'_{:,-j}G_jG_j^0) - G_j y_j \right\|^2 \\ &= \left\| (G_j y_j^0 + a^{-1}Z_{:,-j}\widehat{R}_{Z,:,-j} - a^{-1}Z_{:,-j}Z'_{:,-j}G_j y_j^0) - G_j y_j \right\|^2 \\ &= \left\| z - G_j y_j \right\|^2 \end{aligned}$$

subject to the constraints that $G_j y_j$ is centered and $y'_j G'_j G_j y_j = 1$. This is now a normalized monotone regression problem.[12]

[12]Given weights $w_1, \ldots, w_T$ and real numbers $x_1, \ldots, x_T$, the monotone (isotonic) regression problem finds $\hat{x}_1, \ldots \hat{x}_T$ to minimize $S(y) = \sum_{t=1}^{T} w_t(x_t - y_t)^2$ subject to the monotonicity condition $t \preceq k$ implies $y_t \leq y_k$ where $\preceq$ is a partial ordering on the index set $[1, \ldots T]$. An up-and-down-block algorithm is given in Kruskal (1964). See also de Leeuw (2005).

ACKNOWLEDGMENTS

I thank Aman Ullah for teaching me econometrics and especially grateful for his guidance and support over the years. Comments from two anonymous referees are greatly appreciated. I also thank Nickolay Trendafilov for helpful comments and discussions.

FUNDING

Financial support from the National Science Foundation (SES-0962431) is gratefully acknowledged.

REFERENCES

Almund, M., Duckworth, A., Heckman, J., Kautz, T. (2011). Personality psychology and economics, NBER working paper 16822.

Anderson, T. W., Rubin, H. (1956). Statistical inference in factor analysis. In: Neyman, J., ed. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*. Vol. V. Berkeley: University of California Press, pp. 114–150.

Babakus, E., Ferguson, C., Joreskog, K. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributonal assumptions. *Journal of Marketing Research* 24(2):222–228.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* 71(1):135–172.

Bai, J., Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* 70(1):191–221.

Bai, J., Ng, S. (2008). Large dimensional factor analysis. *Foundations and Trends in Econometrics* 3(2):89–163.

Boivin, J., Ng, S. (2006). Are more data always better for factor analysis. *Journal of Econometrics* 132:169–194.

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

Browne, M. W. (1984). Asymptotically distribution free methods in the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology* 37:62–83.

Buja, A. (1990). Remarks on functional canonical variates: Alternating least squares methods and ACE. *Annals of Statistics* 18(3):1032–1069.

Chamberlain, G., Rothschild, M. (1983). Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica* 51:1281–2304.

Cliff, N. (1966). Orthogonal rotation to congruence. *Psychometrika* 31:33–42.

Connor, G., Korajzcyk, R. (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics* 15:373–394.

Conti, G., Heckman, J., Lopes, H., Piatek, R. (2012). Constructing economically justified aggregates: An application to early orgins of health. *Journal of Econometrics*.

Cunha, F., Heckman, J. (2008). Formulating, identifying and estimating the techology of cognitive and noncognitive skill formuation. *Journal of Human Resources* 43(4):738–782.

de Leeuw, J. (2004). Least squares optimal scaling of partially observed linear systems. In: *Recent Developments on Structural Equation Models*. Kluwer Academic Publishers, pp. 121–134.

de Leeuw, J. (2005). Monotonic regression. In: Everitt, B., Howell, D., ed. *Encyclopedia of Statistics in Behavioral Science*. Vol. 3. New York: John Wiley and Sons, pp. 1260–1261.

Dolan, C. (1994). Factor analysis of variables with 2,3,5, and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology* 46:309–326.

Doz, C., Giannone, D., Reichlin, L. (2007). A Quasi-Maximum Likelihood Approach for Large Approximate Dynamic Factor Models, ECARES working paper.
Filmer, D., Pritchett, L. (1998). Estimating Wealth Effect Without Expenditure Data – Or Tears: An Application to Educational Enrollments in States of India, Discussion Paper 1994, World Bank.
Gower, J., Dijksterhuis, G. (2004). *Procrustes Problems*. Oxford: Oxford University Press.
Harman, H., Jones, W. (1966). Factor analysis by minimizing residuals (Minres). *Psychometrika* 31(3):315–368.
Israels, A. (1984). Redundancy analysis for qualitative variables. *Psychometrika* 49(3):331–346.
Joreskog, K. (1994). On the esetimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika* 59:381–389.
Joreskog, K., Moustki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research* 36:347–387.
Joreskog, K., Sorbom, S. D. (2006). *LISREL User's Reference Guide*, Chicago.
Joreskog, K. G. (1970). A genearl method for analysis of covariance structurs. *Biometrika* 57:239–51.
Joreskog, K. G. (2003). Factor Analysis by MINRES, www.ssicentral.com/lisrel/techdocs/minres.pdf. Last accessed 31 October 2014.
Keller, W., Wansbeek, T. (1983). Multivariate methods for quantitative and qualitative data. *Journal of Econometrics* 22:91–111.
Kiers, H. (1990). Majorization as a tool for optimizing a class of matrix functions. *Psyhometrika* 55(3):417–428.
Kiers, H. (2002). Setting up alternating least squares and iterative majorization algorithms for solving various matrix optimization problems. *Computational Statistics and Data Analysis* 41:157–170.
Kiers, H., Takane, Y., Mooijaart, A. (1993). A mononotically convergent algorithm for FACTALS. *Psychometrika* 58(4):567–574.
Kolenikov, S., Angeles, G. (2009). Socioeconomic status measurement with discrete proxy variables: Is principal component analysis a reliable answer. *Review of Income and Wealth* 55:128–165.
Kruskal, J. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika* 29:115–129.
Kruskal, J., Shepard, R. (1974). A nonmetric variety of linear factor analysis. *Psychometrika* 39:123–157.
Lancaster, H. (1957). Some properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika* 44(1/2):289–292.
Lawley, D. N., Maxwell, A. E. (1971). *Factor Analysis in a Statistical Method*. London: Butterworth.
McDonald, R. P. (1985). *Factor Analysis and Related Methods*. Hilsdale, NJ: Erlbaum.
Meulman, J., Heiser, W. (2001). *SPSS Categories*. 11.0 edn. SPSS Inc.
Michailidis, G., de Leeuw, J. (1998). The gifi system of descriptive multivariate analysis. *Statistical Science* 13(4):307–336.
Muthén, B. (1984). A general structural equation model with dichotomous, ordered, categorical, and continuous latent variable indicators. *Psychometrika* 49:115–132.
Muthen, B., Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Psychology* 38:71–189.
Nevels, K. (1989). An improved solution for factals: A nonmetric common factor analysis. *Psychometrika* 54(3390343).
Olsson, U., Drasgow, F., Dorans, N. (1982). The polyserial correlation coefficient. *Psychometrika* 47(3):337–347.
Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *Review of Economics and Statistics* 92(4):1004–1016.
Pearson, K. (1900). On the correlation of charcters not quantitatively measurable. In: *Mathematical Contributions to the Theory of Evolution: Philosophical Transactions of the Royal Society of London, Series A*. Vol. 195, pp. 1–46.
Racine, J., Li, Q. (2004). Nonparametric estimation of regression functions with both discrete and continuous data. *Journal of Econometrics* 119:99–130.
Schonmenn, P. (1966). A generalized solution of the orthogonal procustes problem. *Psychometrika* 31(1):1–10.
Socan, G. (2003). The Incremental Value of Minimum Rank Factor Analysis. Ph.D. dissertation, University of Groningen.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology* 15:201–293.

Su, L., Ullah, A. (2009). Functional coefficient estimation with both categorical and continuous data. In: Li, Q., Racine, J., ed. *Advances in Econometrics*. Vol. 25. Emerald Group Publishing Limited, pp. 131–167.

Takane, Y., Young, F., de Leeuw, J. (1979). Nonmetric common factor analysis an alternating least squares method with optimal scaling features. *Behaviormetrika* 6:45–56.

ten Berge, J. (1993). *Least Squares Optimization in Multivariate Analysis*. Leiden: DSWO Press.

ten Berge, J., Nevels, K. (1997). A general solution to mosier's oblque procrustes problem. *Psychometrika* 42:593–600.

Tenenhaus, M., Young, F. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika* 50(1):91–119.

Trendafilov, N., Unkel, S. (2011). Exploratory factor analysis of data matrices with more variables than observations. *Journal of Computational Graphical Statistics* 20. Forthcoming.

Unkel, S., Trendafilov, N. (2010). Simultaneous parameter estimation in exploratory factor analysis: An expository review. *International Statistical Review* 78(3):363–382.