Double Instrumental Variable Estimation of Interaction Models with Big Data

Patrick Gagliardini *

Christian Gouriéroux †

First version: December 2014

This version: May 2015 ‡

*Università della Svizzera Italiana (USI), Lugano, and SFI.

†CREST and University of Toronto.

‡Acknowledgements: We are grateful to J.-C. Heam, E. Renault, A. Trognon and the participants at the 6.th French Econometrics Conference (2014) in Paris for useful comments.

Double Instrumental Variable Estimation of Interaction Models with Big Data

Abstract

The static factor analysis of a (n, m) matrix of observations Y is usually based on the joint spectral decomposition of the matrix squares YY' and $Y'Y$. However, for very large dimensions m and n, this approach has a high level of numerical complexity. Indeed, the number of required computations grows cubically w.r.t. the matrix dimensions, that is, much faster than the number of observations. The big data dimensions can be used to propose new estimation methods with a reasonable numerical complexity. The double instrumental variable (IV) approach uses row instruments and column instruments to estimate consistently the factors up to K^2 parameters, where K is the number of factors. Then, the factor model can be concentrated and the K^2 missing parameters estimated consistently by ordinary least squares. We compare the double IV approach to Principal Component Analysis (PCA). The double IV approach can be used for the analysis of recommender systems and is a new collaborative filtering approach.

Keywords: Interaction, Connectedness, Factor Analysis, Principal Component Analysis, Big Data, Panel Data, Instrumental Variable, Recommender System, Completion Matrix, Collaborative Filtering, Preferences.

1 Introduction

The big data challenge has two prominent features, that are the huge number of data items, but also the possibility to study new economic questions, because of new types of available data. Among the most interesting characteristics of big data sources developed in recent years, these datasets provide detailed information on the interdependencies and interactions between the individual behaviour of economic and social agents.

In this paper we consider the interactions in a homogenous population of individuals. These interactions are usually represented by matrices, whose generic element of index (i, j) measures the magnitude of the interaction from individual i to individual j. For instance, the element can be the number of e-mails sent by i to j during a given period: in this case, i is the index of the transmitter and j the index of the receiver. [1] Another example concerns the diffusion of systemic risk in a financial sector. The interconnections are summarized by the exposure matrices available for each class of assets [see e.g. Upper, Worms (2004), Gourieroux, Heam, Monfort (2012)]. The element of the matrix can be the amount of debt (resp. stocks, options) of financial institution i held by institution j: here, i is the index of the debt issuer, whereas j is the index of the debt holder. Similar examples are the observations of the traded volumes between a set of buyers and a set of sellers [Kranton, Minehart (2011)], the co-citations between researchers in Economics, the table of import/export to major trading partners [see e.g. Leng, Tang (2012)], the degree of assistance between individuals measured for instance by money transfers. The indices i and j can have different interpretations, for instance the consumption of good j by household i during a given period of time, or the scores attributed by a list of people to a set of items (movies, books, ...) used to build recommender systems [see e.g. Su, Khashgoftaar (2009)]. Sometimes, the observed matrices are symmetric, for instance when they measure the social distance between individuals with social interactions such as friendship, acquaintance, collaboration [Wasserman, Faust (1994), Nowicki, Snijders (2001), Jackson (2008), Iijima, Kamada (2010)].

The interactions are usually modeled by factor analysis and the factors estimated by standard methods such as the Singular Value Decomposition (SVD), the Principal Component Analysis (PCA), or other reduction techniques. [2] However, these estimation techniques require a number of

[1] Typically the financial supervisory authorities have such information for traders.

[2] See Traxillo (2003) and Suhr (2009) for a description and comparison of the softwares for PCA and Exploratory

computations much larger than the number of data (see the discussion in Section 2.7). Their too large numerical complexity makes them inadequate for huge dimensional matrices of interactions.

The aim of our paper is to explain why the large number of data can greatly facilitate the estimation of such nonlinear models.

We consider in Section 2 the static interaction model and explain how it can be easily estimated by applying linear instrumental variables methods based on asymptotic instruments for the row and column factors, respectively. In this respect we extend to matrix-variates the methodology introduced in Granger (1987), or Forni, Reichlin (1996). Such instruments can be constructed by partial averaging of nonlinear transformations of the interaction data. We derive the asymptotic properties of these linear approaches used to estimate the factor model. We show that the approach can also be applied for models with incomplete data. In this respect it provides a new method of collaborative filtering. Finally, we compare the asymptotic properties of the double IV approach and of Principal Component Analysis. The approach is extended in Section 3 to time series of interaction matrices, that is, to triply indexed observations. In Section 4 we illustrate the double IV estimation technique by a simulation study. Section 5 concludes. The proofs are gathered in Appendices.

2 Static Factor Analysis

2.1 The static interaction model

We consider two populations of individuals indexed by i and j, with $i = 1, \ldots, n$ and $j = 1, \ldots, m$, respectively. We denote $y_{i,j}$ the magnitude of the interaction from i to j. [3]

When these populations and interactions are homogenous, the static model can be written as:

$$y_{i,j} = \alpha_i' \beta_j + \varepsilon_{i,j}, \quad i = 1, \ldots, n, \ j = 1, \ldots, m, \tag{2.1}$$

where α_i and β_j are K-dimensional factors and $\varepsilon_{i,j}$ is a one-dimensional error term. The homogeneity assumption is:

Factor Analysis available in SAS.

[3] Alternatively, we have one population of individuals i, and a set of items j. Then, $y_{i,j}$ denotes either the consumption of item j by individual i, or the opinion of individual i on item j.

Assumption A.1: Homogeneity

All variables α_i, β_j, $\varepsilon_{i,j}$ are independent. The $\alpha_i's$ (resp. the $\beta_j's$, the $\varepsilon_{i,j}'s$) are identically distributed, with finite second-order moments.

Under Assumption A.1, the factor model treats in a symmetric way the stochastic factors associated with individual i and individual j.

We will also use, when necessary, the zero-mean assumption.

Assumption A.2: Zero-mean

The variables α_i, β_j and $\varepsilon_{i,j}$ have zero-mean.

Factor model (2.1) can be written in matrix notation as:

$$Y = \alpha\beta' + \varepsilon, \tag{2.2}$$

where $Y = (y_{i,j})$ is the (n, m) matrix of observations, α (resp. β) the (n, K) [resp. (m, K)] matrix of factor observations, and ε the (n, m) matrix of error terms. For a given matrix such as Y, we denote y_i the $(m, 1)$ vector $y_i = (y_{i,j}, j = 1, \ldots, m)$, that is the transposed of row i of matrix Y, and by y^j its j-th n-dimensional column vector.

Under Assumption A.1 (resp. Assumptions A.1-A.2), the factors α_i and β_j are identifiable up to an invertible linear transformation. In other words, we can identify the vector spaces spanned by the latent factors, but not the factor values themselves.

Model (2.1) reduces the dimensionality of the distributional problem. Indeed, the nm-dimensional distribution of matrix-variate Y is characterized by the two K-dimensional distributions of the $\alpha's$ and $\beta's$ plus the one-dimensional distribution of the $\varepsilon's$. Model (2.1) introduces pairwise dependence between the elements of matrix Y through rows and columns. This dependence is not visible when we only consider second-order moments (when they exist), since:

$$\begin{aligned} Cov(y_{i,j}, y_{k,l}) &= Cov(\alpha_i'\beta_j, \alpha_k'\beta_l) \\ &= Cov\{E(\alpha_i'\beta_j|\beta), E(\alpha_k'\beta_l|\beta)\} + E\{Cov(\alpha_i'\beta_j, \alpha_k'\beta_l|\beta)\} \\ &= 0, \text{ if } i \neq k, \end{aligned}$$

from Assumptions A.1-A.2. By symmetry we deduce that all pairs of elements of matrix Y are marginally uncorrelated. However, the observations associated with two different dyads are not necessarily independent as for instance they are in the model introduced in Holland, Leinhardt (1981) for binary relations.

In fact, model (2.1) satisfies the transitivity condition, which is often mentioned as an important feature of social networks. [4] Indeed, the magnitude of the link between dyads is larger if they have an actor in common. This is a form of spatial Markov dependence [see e.g. Frank, Strauss (1986)]. More precisely, let us consider the case $K = 1$ for expository purpose. If $i \neq k$ and $j \neq l$, the two variables $y_{i,j}$ and $y_{k,l}$ are independent. Let us now consider two dyads with a common actor, that are (i, j) and (k, j) with $i \neq k$, say. We have:

$$\begin{aligned} P[y_{i,j} \in A, y_{k,j} \in B] &= E\{P[\alpha_i\beta_j + \varepsilon_{i,j} \in A|\beta_j]P[\alpha_k\beta_j + \varepsilon_{k,j} \in B|\beta_j]\} \\ &\quad \text{(by the independence of } y_{i,j} \text{ and } y_{k,j} \text{ conditional on } \beta_j\text{)} \\ &\neq E\{P[\alpha_i\beta_j + \varepsilon_{i,j} \in A|\beta_j]\}E\{P[\alpha_k\beta_j + \varepsilon_{k,j} \in B|\beta_j]\} \\ &= P[y_{i,j} \in A]P[y_{k,j} \in B], \end{aligned}$$

for Borel sets A and B. The two dyads are not independent, and the dependence can be either positive, or negative. Therefore, model (2.1) is very different from the matrix-variate normal models with a constrained variance-covariance matrix for the elements of Y [see e.g. Dawid (1981), Gupta, Nagar (2000), or Leng, Tang (2012)].

2.2 The double instrumental variable approach

Let us now explain how to estimate consistently factors α_i, β_j and error terms $\varepsilon_{i,j}$ (and also their distributions), when both dimensions n and m tend to infinity. Our approach relies on the use of instrumental variables [Theil (1953)]. We consider a model satisfying Assumptions A.1-A.2, and start with the case of a minimal number of instruments (just-identified setting). [5] We denote by Rk the rank of a matrix.

[4] The two other important features of a social network are homophily on unobserved attributes and clustering [see the discussion in Handcock, Raftery and Tantrum (2007)]. The homophily on unobserved attributes is introduced in Appendix 2, and clustering is discussed in Section 2.5.2.

[5] The overidentified case is discussed in Section 2.6.

Definition 1: *The variables x_i, with $i = 1, \ldots, n$ (resp. z_j, with $j = 1, \ldots, m$) is a minimal set of instrumental variables for factor β_j (resp. factor α_i) if and only if:*

i) $\dim x_i = K$ *[resp.* $\dim z_j = K$*].*

ii) $\underset{n\to\infty}{\text{plim}} \frac{1}{n} \sum_{i=1}^{n} x_i \alpha_i' \equiv C(x, \alpha)$*, say, where* $Rk[C(x, \alpha)] = K$;

[resp. $\underset{m\to\infty}{\text{plim}} \frac{1}{m} \sum_{j=1}^{m} z_j \beta_j' \equiv C^*(z, \beta)$*, say, where* $Rk[C^*(z, \beta)] = K$*].*

iii) $\underset{n\to\infty}{\text{plim}} \frac{1}{n} \sum_{i=1}^{n} x_i \varepsilon_{i,j} = 0, \forall j$*; [resp.* $\underset{m\to\infty}{\text{plim}} \frac{1}{m} \sum_{j=1}^{m} z_j \varepsilon_{i,j} = 0, \forall i$*].*

The stochastic convergence conditions in Definition 1 are in particular satisfied under the following assumption.

Assumption A.3: Conditions on the instruments

i) The instruments $x_i, i = 1, \ldots, n$ [resp. $z_j, j = 1, \ldots, m$] are such that the pairs (x_i, α_i) [resp. (z_j, β_j)] are i.i.d. with finite second-order moments.

ii) The pairs (x_i, α_i) and (z_j, β_j) are independent. They are also independent of the errors $\varepsilon_{i,j}$.

Under Assumption A.3, the limits in Definition 1 are:

$$C(x, \alpha) = E(x_i \alpha_i'), \quad C^*(z, \beta) = E(z_j \beta_j').$$

Let us denote X (resp. Z) the (n, K) [resp. (m, K)] matrix of observations of the instrumental variables for factor β_j (resp. factor α_i). Under Assumption A.3 the conditions in Definition 1 imply (see Appendix 1.1):

$$\underset{n\to\infty}{\text{plim}} \frac{1}{n} X'Y = C(x, \alpha)\beta' = E(x_i \alpha_i')\beta', \tag{2.3}$$

$$\underset{n\to\infty}{\text{plim}} \frac{1}{m} Z'Y' = C^*(z, \beta)\alpha' = E(z_j \beta_j')\alpha'. \tag{2.4}$$

Moreover, by the identification condition, we can define factors α and β such that:

$$C^*(z, \beta) = E(z_j \beta_j') = Id_K. \tag{2.5}$$

Such K^2 identification restrictions are very appropriate in our instrumental variable framework. They differ from the restrictions introduced in the standard softwares, or considered in the academic literature on factor models [see e.g. Bai, Ng (2013)]. Thus, we directly deduce from (2.3)-(2.4) consistent approximations of α and β as:

$$\begin{cases} \hat{\alpha} &= \frac{1}{m} YZ, \\ \\ \tilde{\beta} &= \frac{1}{n} Y'XC', \end{cases} \tag{2.6}$$

where $C = C(x, \alpha)^{-1}$ is an unknown (K, K) matrix.

Let us now substitute the expressions (2.6) in equation (2.2). We get:

$$Y \simeq \frac{1}{m} YZC \frac{1}{n} X'Y + \varepsilon. \tag{2.7}$$

We get a model which is asymptotically linear w.r.t. the unknown matrix C. Since:

$$Vec\,(ABC) \;= (C' \otimes A)\; vecB, \tag{2.8}$$

where $\otimes$ denotes the Kronecker product [see e.g. Magnus, Neudecker (1994)], we can vectorize matrix system (2.7) to get:

$$\begin{aligned} vec\, Y &\simeq \hat{D}\; vec\, C + vec\, \varepsilon, \\ \text{where } \hat{D} &= \left(\frac{1}{n} Y'X\right) \otimes \left(\frac{1}{m} YZ\right). \end{aligned} \tag{2.9}$$

We deduce consistent estimators of $C, \alpha_i, \beta_j, \varepsilon_{i,j}$ by performing the appropriate regressions.

Proposition 1: *Under Assumptions A.1-A.3, and if X and Z are minimal sets of instrumental variables for β and α, respectively, we have:*

i) *$\hat{C}$ defined by:*

$$vec\, \hat{C} = (\hat{D}'\hat{D})^{-1} \hat{D}'\; vec\, Y,$$

is a consistent estimator of C, when $n, m \to \infty$.

ii) *$\hat{\alpha}_i = \frac{1}{m}(YZ)_i$ is a consistent "estimator" of α_i, for any i.*

iii) *$\hat{\beta}_j = \frac{1}{n}(Y'X\hat{C}')_j$ is a consistent "estimator" of β_j, for any j.*

iv) $\hat{\varepsilon}_{ij} = y_{i,j} - \hat{\alpha}_i' \hat{\beta}_j$ *is a consistent "estimator" of* $\varepsilon_{i,j}$*, for any pair* (i, j).

Proof: See Appendix 1.3 for the proof of the consistency of $\hat{C}$. The other consistency properties immediately follow.

QED

The expression of the double IV estimator $\hat{C}$ can be written under a matrix form as (see Appendix 1.2):

$$\hat{C} = (\frac{1}{m^2} Z'Y'YZ)^{-1} (\frac{1}{m} Z'Y') Y (\frac{1}{n} Y'X)(\frac{1}{n^2} X'YY'X)^{-1}. \tag{2.10}$$

Thus, we get consistent estimators of the $\alpha_i, \beta_j, \varepsilon_{i,j}$ with simple closed form expressions which require only the inversion of matrices with the reasonable dimension (K, K). The computational complexity of the double IV estimator, that is the number of operations necessary to compute the estimates, is $O(nm)$, i.e., the order of the sample size. Indeed, computing matrices YZ and $Y'X$ requires $O(nm)$ operations, and the same holds for matrices $\hat{C}$, $\hat{\alpha}$, $\hat{\beta}$ from equation (2.10) and Proposition 1.

Note that the identification restriction (2.5) involves the instruments. More precisely, if the instruments Z are replaced by equivalent instruments ZA, say, where A is an invertible (K, K) matrix, the factor α is changed into αA. Thus the interpretation of the factors is modified, but the vector spaces spanned by these factors stay the same as well as the estimated vector spaces. This is summarized in the following Corollary:

Corollary 1: *The double IV estimator of the vector space spanned by the* α_i *(resp.* β_j*) is invariant by a one-to-one linear change of instruments* Z *(resp.* X*).*

The estimated $\alpha's$ and $\beta's$ can be used to construct measures of similarity between individuals i (resp. items j). These measures have to be independent of the selected representer of α, that is, invariant with respect to a linear one-to-one transformation $\alpha \to \alpha Q$, say. Such a measure between individuals i_1 and i_2 is:

$$d(i_1, i_2) = (\hat{\alpha}_{i_1} - \hat{\alpha}_{i_2})' (\frac{1}{n} \sum_{i=1}^{n} \hat{\alpha}_i \hat{\alpha}_i')^{-1} (\hat{\alpha}_{i_1} - \hat{\alpha}_{i_2}). \tag{2.11}$$

These similarity measures could be used to construct a segmentation of the population of individuals.

2.3 Extensions

In this section we show that the double IV approach is easily extended to factor models where the factors do not have zero-mean, or to models including observable covariates. The double IV approach can also be used for estimation of more structural versions of model (2.1), which describe social distances between individuals (see Appendix 2). More important, we explain below how (asymptotic) instruments can be constructed by considering row or column averages of nonlinear transformations of the endogenous observations $y_{i,j}$.

i) General Case

Let us consider the factor model without the zero-mean assumption A.2. We have:

$$y_{i,j} = \alpha_i'\beta_j + \varepsilon_{i,j}.$$

We get a special two-way analysis of variance (ANOVA) model:

$$\begin{aligned} y_{i,j} &= [E(\alpha_i)'E(\beta_j) + E(\varepsilon_{i,j})] + E(\alpha_i)'[\beta_j - E(\beta_j)] \\ &\quad +[\alpha_i - E(\alpha_i)]'E(\beta_j) + [\alpha_i - E(\alpha_i)]'[\beta_j - E(\beta_j)] + [\varepsilon_{i,j} - E(\varepsilon_{i,j})], \end{aligned}$$

with constrained interaction term and links between the marginal and cross-effects. This specification avoids the overparametrization encountered in unconstrained ANOVA [see e.g. Davies (2012)].

Let us now apply the standard ANOVA to matrix Y, that is, replace Y by $\tilde{Y} = (\tilde{y}_{i,j})$, where:

$$\tilde{y}_{i,j} = y_{i,j} - y_{i,\cdot} - y_{\cdot,j} + y_{\cdot,\cdot}, \tag{2.12}$$

where $y_{i,\cdot} = \dfrac{1}{m}\sum_{j=1}^{m} y_{i,j}$, $y_{\cdot,j} = \dfrac{1}{n}\sum_{i=1}^{n} y_{i,j}$, and $y_{\cdot,\cdot} = \dfrac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{i,j}$. For n and m large, we get:

$$\begin{aligned} y_{i,\cdot} &\simeq \alpha_i'E(\beta_j) + E(\varepsilon_{i,j}), \\ y_{\cdot,j} &\simeq E(\alpha_i')\beta_j + E(\varepsilon_{i,j}), \quad y_{\cdot,\cdot} \simeq E(\alpha_i')E(\beta_j) + E(\varepsilon_{i,j}). \end{aligned} \tag{2.13}$$

We deduce:

$$\begin{aligned}\tilde{y}_{i,j} &\simeq \alpha_i'\beta_j - [\alpha_i' E(\beta_j) + E(\varepsilon_{i,j})] - [E(\alpha_i')\beta_j + E(\varepsilon_{i,j})] + E(\alpha_i')E(\beta_j) + E(\varepsilon_{ij}) + \varepsilon_{ij} \\ &= (\alpha_i - E\alpha_i)'(\beta_j - E\beta_j) + \varepsilon_{i,j} - E(\varepsilon_{i,j}). \qquad (2.14)\end{aligned}$$

We get the following Corollary:

Corollary 2: *Under Assumptions A.1 and A.3 and if X and Z are minimal sets of instrumental variables for β and α, respectively, simple consistent estimators of C, $\alpha_i^* = \alpha_i - E\alpha_i$, $\beta_j^* = \beta_j - E\beta_j$ and $\varepsilon_{i,j}^* = \varepsilon_{i,j} - E(\varepsilon_{i,j})$ are obtained by applying the formulas of Proposition 1 after replacing Y by $\tilde{Y}$.*

Then, the drifts $E(\alpha_i), E(\beta_j), E(\varepsilon_{i,j})$ are deduced from the system of equations (2.13). Indeed, we have:

$$\begin{aligned}y_{i,\cdot} &\simeq \hat{\alpha}_i^{*'} E(\beta_j) + E(\alpha_i)'E(\beta_j) + E(\varepsilon_{i,j}) \\ &\simeq \hat{\alpha}_i^{*'} E(\beta_j) + y_{\cdot,\cdot}.\end{aligned}$$

Thus, $E(\beta_j)$ can be estimated consistently by regressing the averages $y_{i,\cdot}$ on the estimated $\hat{\alpha}_i^*$ across i. The expectation $E(\alpha_i)$ is deduced symmetrically. Finally, a consistent estimator of $E(\varepsilon_{i,j})$ is $y_{\cdot,\cdot} - \hat{E}(\alpha_i)'\hat{E}(\beta_j)$.

ii) Factor augmented regression

The factors can also be introduced in a regression model with observable covariates [see e.g. Bai, Ng (2008)]. The model becomes:

$$y_{i,j} = w_{i,j}'a + \alpha_i'\beta_j + \varepsilon_{i,j}, \qquad (2.15)$$

where $w_{i,j}$ is a vector of L observed explanatory variables, which may be correlated with the factors α_i and β_j, but are independent of the error term. The double IV method is easily adjusted to account for such observed variables. By using the instruments X and Z, the analogues of equations (2.3)

are:

$$\left\{\begin{array}{lcl} \underset{n\to\infty}{plim}\frac{1}{n}X'Y & = & \underset{n\to\infty}{plim}\frac{1}{n}X'(\sum_{l=1}^{L}W_l a_l) + C(x,\alpha)\beta', \\ \\ \underset{m\to\infty}{plim}\frac{1}{m}Z'Y' & = & \underset{m\to\infty}{plim}\frac{1}{m}Z'(\sum_{l=1}^{L}W_l' a_l) + C^*(z,\beta)\alpha', \end{array}\right. \tag{2.16}$$

where W_l is the matrix of observations of the l^{th} explanatory variable. Under the identification restriction (2.4), we deduce :

$$\left\{\begin{array}{lcl} \hat{\alpha} & = & \frac{1}{m}YZ - \frac{1}{m}(\sum_{l=1}^{L}W_l a_l)Z, \\ \\ \tilde{\beta} & = & \frac{1}{n}Y'XC' - \frac{1}{n}(\sum_{l=1}^{L}W_l a_l)'XC'. \end{array}\right.$$

The expressions above can be introduced in system (2.15) to get a regression model which is linear in parameter C and quadratic in parameter a. Then, this regression model is easily estimated by nonlinear least squares (NLLS), which is applied to a number of unknown parameters $K^2 + L$ independent of n and m.

An even simpler estimation approach can be introduced under the following additional assumption:

Assumption A.4 : *The w_{ij}'s, (α_i, x_i)'s, (β_j, z_j)'s and ε_{ij}'s are mutually independent.*

Under Assumption A.4, we get a consistent estimator of parameter a by regressing the endogenous variables $y_{i,j}$ on the explanatory variables $w_{i,j}$. Let us denote by $\hat{a}$ this OLS estimator and by $\hat{v}_{i,j} = y_{i,j} - w_{i,j}'\hat{a}$ the residuals in this regression. We have the following Corollary:

Corollary 3: *Under Assumptions A.1, A.2 and A.4:*
i) *the coefficient a of the explanatory variables is consistently estimated by regressing the $y_{i,j}$'s on the $w_{i,j}$'s.*
ii) *Then, the factors and the error terms are consistently estimated by replacing the matrix Y by the matrix of residuals $\hat{V} = (\hat{v}_{i,j})$ in the formulas of Proposition 1.*

Corollary 3 corresponds to the Frisch-Waugh theorem in our setting with observable explanatory variables and latent factors.

iii) Asymptotic instrumental variables

The consistency results are also valid for asymptotic instruments $\hat{x}_i$, say, which may depend on n and m, and tend to the true instruments x_i, when n, m tend to infinity. We get the next corollary:

Corollary 4: *Under Assumptions A.1-A.3 and if $\hat{X}$, $\hat{Z}$ are minimal sets of asymptotic instrumental variables for β and α, respectively, that is, if $plim\ \hat{x}_i = x_i, \forall i,\ plim\ \hat{z}_j = z_j, \forall j$, we get simple consistent estimators of C, α_i, β_j, $\varepsilon_{i,j}$ by applying the formulas of Proposition 1 after replacing X, Z by $\hat{X}$, $\hat{Z}$.*

Let us now show how to easily derive asymptotic instrumental variables for either α, or β by appropriate averaging of transformations of the interaction data.

Proposition 2: *Let us consider a (possibly) nonlinear K-dimensional mapping c and define:*

$$\frac{1}{m}\sum_{j=1}^{m} c(y_{i,j}) \equiv c_{i,\cdot}, \quad \frac{1}{n}\sum_{i=1}^{n} c(y_{i,j}) \equiv c_{\cdot,j}.$$

In general, $c_{i,\cdot}$ (resp. $c_{\cdot,j}$) can be used as an approximate instrumental variable $\hat{x}_i$ for β_j (resp. $\hat{z}_j$ for α_i).

Proof: Under Assumptions A.1-A.2, we have:

$$\begin{aligned} c_{i,\cdot} &= \frac{1}{m}\sum_{j=1}^{m} c(y_{i,j}) = \frac{1}{m}\sum_{j=1}^{m} c(\alpha_i'\beta_j + \varepsilon_{i,j}) \\ &\simeq \int\int c(\alpha_i'\beta_j + \varepsilon_{i,j}) dG_\beta(\beta_j) dG_\varepsilon(\varepsilon_{i,j}), \end{aligned}$$

as $m \to \infty$, where G_β and G_ε denote the (true) distributions of β_j and $\varepsilon_{i,j}$, respectively. Therefore $\hat{x}_i \equiv c_{i,\cdot}$ tends to a deterministic (unknown) function of α_i, that is $x_i = a(\alpha_i)$, say. Similarly, $c_{\cdot,j}$ tends to a deterministic function of β_j, that is $z_j = b(\beta_j)$, say. Thus, the conditions in Assumption A.3 are satisfied asymptotically with $E(x_i\alpha_i') = E[a(\alpha_i)\alpha_i']$ and $E(z_j\beta_j') = E[b(\beta_j)\beta_j']$. In this framework the identification restriction becomes :

$$E[b(\beta_j)\beta_j'] = Id_K.$$

The asymptotic instruments are valid whenever $Rk\ E[a(\alpha_i)\alpha_i'] = K$.

QED

iv) Missing data and collaborative filtering

The recommender systems [6] collect the ratings posted by n users on m different items: books, movies, cosmetics, etc. However, these observations are sparse, since a given user has experimented a limited number of items. Two questions arise in such an incomplete data framework: a) How to estimate the underlying parameters α_i, β_j, ... ? b) How to complete for missing data, a question known as matrix completion or collaborative filtering [7] ?

We explain below why the double IV approach answers these questions by applying the computation on observed data (ratings) only. More precisely, let us consider the following extension of model (2.1) :

$$y_{i,j} = (\alpha_i'\beta_j + \varepsilon_{i,j})\xi_{i,j}, \tag{2.17}$$

where the variables $\xi_{i,j}$ are indicator variables assumed i.i.d. and independent of the $\alpha_i's$, $\beta_j's$ and $\varepsilon_{i,j}'s$. The indicator value is $\xi_{i,j} = 1$, if the rating is posted, and $\xi_{i,j} = 0$, otherwise.

The modelling in equation (2.17) assumes implicitly that the users have similar behavior and rate the items similarly. Moreover, the assumptions on indicator variables ξ imply that there is no endogenous selectivity in the decision of posting a rating. These assumptions are standard in the literature [see e.g. Klopp (2012), eq. 1]. Of course, the no selectivity assumption is not satisfied, if some users decide to rate only the items that they like (resp. they don't like). The model also assumes continuous scores (ratings).

To understand why the double IV approach still works in this framework with missing data, let

[6] developed especially for large online companies like eBay, Amazon, Expedia, Pandora. See also the famous Netflix problem [ACM, SIGKDD and Netflix (2007)].

[7] The term "collaborative filtering" has been first introduced in Goldberg et al. (1992).

us consider the appropriate averages computed on observed data only. We have:

$$\frac{\sum_{i=1}^{n}[x_i(\alpha_i'\beta_j+\varepsilon_{i,j})\xi_{i,j}]}{\sum_{i=1}^{n}\xi_{i,j}}$$

$$\simeq \frac{E(x_i\alpha_i'\xi_{i,j})\beta_j+E(x_i\varepsilon_{i,j}\xi_{i,j})}{E(\xi_{i,j})}, \text{ where the expectation is w.r.t. the user uncertainty,}$$

$$= \frac{E(x_i\alpha_i')\beta_j E(\xi_{i,j})+E(x_i\varepsilon_{ij})E(\xi_{i,j})}{E(\xi_{i,j})}, \text{ due to the assumption of independence on } \xi,$$

$$= C(x,\alpha)\beta_j,$$

which is the analogue of (2.3). Moreover, when the zero-mean condition in Assumption A.2 is not satisfied, the averaging is applied after the ANOVA transformation:

$$\begin{aligned}\tilde{y}_{i,j} &= y_{i,j}-(y_{i,\cdot}/\xi_{i,\cdot})\xi_{i,j}-(y_{\cdot,j}/\xi_{\cdot,j})\xi_{i,j}+(y_{\cdot,\cdot}/\xi_{\cdot,\cdot})\xi_{i,j}\\ &\simeq \xi_{i,j}\left\{[\alpha_i-E(\alpha_i)]'[\beta_j-E(\beta_j)]+\varepsilon_{i,j}-E(\varepsilon_{i,j})\right\}. \qquad (2.18)\end{aligned}$$

Similarly let us consider the implementation of the asymptotic instrumental variable as in the previous subsection. We get:

$$\frac{\sum_{j=1}^{m}c(y_{i,j})\xi_{i,j}}{\sum_{j=1}^{m}\xi_{i,j}} = \frac{\sum_{j=1}^{m}c(\alpha_i'\beta_j+\varepsilon_{i,j})\xi_{i,j}}{\sum_{j=1}^{m}\xi_{i,j}}$$

$$\sim \int\int c(\alpha_i'\beta_j+\varepsilon_{i,j})dG_\beta(\beta_j)dG_\varepsilon(\varepsilon_{i,j}),$$

which is the analogue of the result in Proposition 2. Therefore, the double IV approach can be applied whenever $P(\xi_{i,j}=1)>0$. In practice, the method has to be applied only to users and items with a sufficient number of complete data (more than 30, say).

Once the α_i,β_j are estimated, model (2.17) is easily used for prediction (filtering) purpose. An incomplete data (i,j), such that $\xi_{i,j}=0$, will be predicted by $\hat{\alpha}_i'\hat{\beta}_j$. These filtered values can be

used to recommend to a user a list of new preferred items by computing the N top-ranked items, after having ranked the predicted ratings $\hat{\alpha}_i'\hat{\beta}_j$ by decreasing order.

Thus, the double IV approach is a new model-based collaborative filtering methodology useful to analyze the individual preferences of individuals. This is a competitor of alternative methodologies based on latent semantic models [see e.g. Hofmann (2001), (2004)] or on multinomial mixture models [see e.g. Miyahara, Pazzani (2002)].

In the special case where there is no data noise, i.e. $\varepsilon_{i,j} = 0$, $\forall i, j$, the problem consists in reconstructing a matrix of rank K from random sampling of its elements. In this case, the double IV approach is an alternative to the approaches based on either convex optimization, or minimization of the nuclear norm [see e.g. Candes, Recht (2009), Recht (2009)]. This literature looks for the number of randomly selected entries required to reconstruct an unknown low rank matrix. This technique has been extended to noisy entries by considering for instance nuclear norm penalized estimators [see e.g. Klopp (2012)]. Our approach avoids these techniques by using the weak distributional assumptions introduced on the low rank matrix.

2.4 Asymptotic behaviour of the double IV estimator

Since the estimators have simple closed form expressions, it is easy to derive their asymptotic distributions. This derivation is performed in Appendix 1.4. We have the following properties:

Proposition 3: *Let us assume that dimensions n and m tend to infinity at equivalent rates $m = \mu n + o(n), \mu \geq 1$.*

i) *The first-order expansion for estimator $\hat{C}$ is:*

$$\begin{aligned}\sqrt{n}(\hat{C} - C) &= -\frac{1}{\sqrt{\mu}}\{\frac{1}{\sqrt{m}}\sum_{j=1}^{m}[\beta_j z_j' - E(\beta_j z_j')]\}C \\ &\quad -C\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[x_i\alpha_i' - E(x_i\alpha_i')]\}C + o_p(1),\end{aligned}$$

where $C = [E(x_i\alpha_i')]^{-1}$, and $o_p(1)$ denotes a negligible term in probability.

ii) *The estimator $\hat{C}$ is asymptotically normal, with the asymptotic variance-covariance matrix :*

$$V_{as}[vec(\sqrt{n}(\hat{C} - C))] = \frac{1}{\mu}(C' \otimes Id)V[z_j \otimes \beta_j](C \otimes Id) + (C' \otimes C)V[\alpha_i \otimes x_i](C \otimes C').$$

We observe the effect of errors-in-variables on factors α and β. This implies the rate of convergence $1/\sqrt{n}$ for estimator $\hat{C}$, instead of the rate $1/n$ that would apply if the factors were observable.

We have noted that matrix C is not invariant when the basic instruments are replaced by equivalent ones in a one-to-one linear relationship. An invariant of the problem is the product $\alpha\beta'$. Therefore, it is interesting to consider also the expansion of the matrix of fitted values [8] : $\hat{Y} = \hat{\alpha}\hat{\beta}'$.

Proposition 4: *Let us assume that dimensions n and m tend to infinity at equivalent rates $m = \mu n + o(n), \mu \geq 1$.*

i) *The first-order expansion for $\hat{Y} = \hat{\alpha}\hat{\beta}'$ is:*

$$\sqrt{n}(\hat{Y} - \alpha\beta') = \frac{1}{\sqrt{\mu}}(\frac{1}{\sqrt{m}}\varepsilon Z)[E(\beta_j z_j')]^{-1}\beta' + \alpha[E(x_i\alpha_i')]^{-1}(\frac{1}{\sqrt{n}}X'\varepsilon) + o_p(1).$$

ii) *The double IV estimator of $\alpha\beta'$ is asymptotically normal. Its asymptotic variance is given by:*

$$\begin{aligned} V_{as}[vec(\sqrt{n}(\hat{Y} - \alpha\beta'))] &= \frac{\sigma^2}{\mu}\{\beta[E(z_j\beta_j')]^{-1}[E(z_j z_j')][E(\beta_j z_j')]^{-1}\beta'\} \otimes Id_n \\ &\quad + \sigma^2 Id_m \otimes \{\alpha[E(x_i\alpha_i')]^{-1}E(x_i x_i')[E(\alpha_i x_i')]^{-1}\alpha'\}. \end{aligned}$$

The asymptotic variance of $\hat{Y}$ has a simple form, in which we recognize standard asymptotic variances of IV estimators:

$$\sigma^2\{E(\beta_j z_j')E(z_j z_j')^{-1}E(z_j\beta_j')\}^{-1} \text{ and } \sigma^2\{E(\alpha_i x_i')E(x_i x_i')^{-1}E(x_i\alpha_i')\}^{-1}.$$

Indeed, these matrices are the asymptotic variance matrices of exactly identified IV estimators with explanatory variables β_j and instruments z_j (resp. explanatory variables α_i and instruments x_i), and conditionally homoscedastic errors.

The finite sample properties of estimator $\hat{Y}$ can be easily derived by bootstrapping the empirical distributions of $\hat{\alpha}_i$, $\hat{\beta}_j$, $\hat{\varepsilon}_{ij}$, respectively. The numerical complexity of this bootstrap procedure is of order $nm \times S$, where S is the number of replications in the bootstrap.

[8] When dimensions n, m increase, the dimensions of matrix $\hat{Y}$ increase too. For expository purpose, we do not discuss this point. The results in Proposition 4 are valid for any given submatrix of $\hat{Y}$ including the (N, M) first pairs, with $N \leq n$, $M \leq m$, and N, M fixed in the asymptotics.

2.5 Estimation of the unknown distributions

Once the core matrix of parameters C has been estimated, we deduce consistent approximations of the components of row and column factors $\hat{\alpha}_i$ and $\hat{\beta}_j$, and of the errors $\hat{\varepsilon}_{i,j}$ (see Proposition 1). They converge at rate $1/\sqrt{n}$ due to the error-in-variables effect. They can be used to derive consistent parametric or nonparametric estimators of the distributions of α_i, β_j, $\varepsilon_{i,j}$.

Let us first focus on nonparametric inference.

2.5.1 Nonparametric inference

Two cases have to be distinguished.

i) Continuous distributions

If α_i (resp. $\beta_j, \varepsilon_{i,j}$) has a continuous distribution, and if the number K of factors is not too large, the corresponding density can be estimated by using a kernel estimator based on the estimates $\hat{\alpha}_i$ (resp. $\hat{\beta}_j$, $\hat{\varepsilon}_{i,j}$). Since the error on α_i is at a parametric rate, these kernel estimators have standard asymptotic properties. When the data are incomplete, we have mentioned that the estimation of the $\alpha'_j s$ and $\beta'_j s$ applies only for the users and items for which enough rating observations are available. The estimated distributions of the $\alpha' s$ and $\beta' s$ can then be used to predict the unobserved ratings for a user (resp. an item) with few observed responses. We have just to apply a Bayesian updating based on the few observed ratings and the estimated distributions as prior distributions.

ii) Mixed distributions

The situation is different if the distribution is a mixture of continuous distributions and point masses at zero. Indeed, even if some underlying components of α_i are equal to zero, their approximations based on $\hat{\alpha}_i$ are almost surely nonzero. Thus, we have first to define an interval around zero used to assign to zero all the components of $\hat{\alpha}_i$ of interest in this interval, and to the continuous component of the distribution all the values of the components of $\hat{\alpha}_i$ outside this interval. The length of this interval has to account for the accuracy of estimator $\hat{\alpha}_i$. We have the following Proposition :

Proposition 5: *Let $\mathcal{A}$ be a subset of $\{1, \ldots, K\}$ and $P(\alpha, \mathcal{A}) = P[\alpha_{i,k} = 0, k \in \mathcal{A}]$. A consistent*

estimator of probability $P(\alpha, \mathcal{A})$ is:

$$\hat{P}_n(\alpha, \mathcal{A}) = \frac{1}{n}\sum_{i=1}^{n}(\prod_{k\in\mathcal{A}} \mathbb{1}_{|\hat{\alpha}_{i,k}|<h_n}),$$

where h_n tends to zero, when $n \to \infty$, and is such that $\frac{\sqrt{\log\log n}}{\sqrt{n}} = o(h_n)$.

Proof: This is a direct consequence of the Law of Iterated Logarithm.

QED

To understand how this result will be used in practice, let us consider the case $K = 2$, and denote :

$$\mathcal{I}_n(\alpha, \mathcal{A}) = \{i : |\hat{\alpha}_{ik}| < h_n, \forall k \in \mathcal{A}\}.$$

Proposition 5 will be used to estimate the three probabilities $P(\alpha, \mathcal{A})$ by $\hat{P}(\alpha, 1)$, $\hat{P}(\alpha, 2)$, $\hat{P}(\alpha, (1, 2))$ and to determine the associated sets $\mathcal{I}_n(\alpha, 1)$, $\mathcal{I}_n(\alpha, 2)$, $\mathcal{I}_n(\alpha, \{1; 2\})$.

Then the joint distribution of $(\alpha_{i1}, \alpha_{i2})$ has different continuous components:

- a bivariate continuous density, which is estimated by applying a bivariate kernel to the set of observations $\{i : |\hat{\alpha}_{i,1}| > h_n, |\hat{\alpha}_{i,2}| > h_n\}$;
- a one-dimensional continuous density for α_2, when $\alpha_1 = 0$, estimated by applying a one-dimensional kernel to α_2, to the set of observations $\{i : |\hat{\alpha}_{i,1}| < h_n, |\hat{\alpha}_{i,2}| > h_n\}$;
- a one dimensional continuous density for α_1, when $\alpha_2 = 0$, estimated by applying a one-dimensional kernel to α_1, to the set of observations $\{i : |\hat{\alpha}_{i,1}| > h_n, |\hat{\alpha}_{i,2}| < h_n\}$.

2.5.2 Parametric inference and clustering

Parametric inference is especially appealing to assign the individuals to different clusters. This is easily done as follows. First assume that the distribution of α is a mixture of K_α parametric distributions:

$$G_\alpha(.) = \sum_{k=1}^{K_\alpha} \pi_{\alpha,k} G_{\alpha,k}(.; \theta_{\alpha k}),$$

where $\theta_{\alpha k}$ and $\pi_{\alpha,k}, k = 1, \ldots, K$ are unknown parameters, and the same for the distribution of β:

$$G_\beta(.) = \sum_{k=1}^{K_\beta} \pi_{\beta,k} G_{\beta,k}(.; \theta_{\beta k}).$$

Then, a standard method to estimate these mixtures and assign the individuals can be applied to the consistent approximations $\hat{\alpha}_i$ of α_i (resp. $\hat{\beta}_j$ of β_j). Thus, the methodology can be used to construct bidimensional clusters, which are obtained by crossing the clusters of the α_i and the clusters of the β_j.

In this respect this model is an alternative to other model-based analyses such as the stochastic block structure models, in which the blocks (clusters) are also latent and estimated from the data [see e.g. Wasserman, Anderson (1987), Nowicki, Snijders (2001), Handcock, Raftery, Tantrum (2007), Latouche, Birmele, Ambroise (2011)], or the probabilistic latent semantic analysis [Hofmann (2003)].

2.6 The generalized double IV estimator

From Proposition 1 and equations (2.6) and (2.10), we see that the matrix of fitted values is given by:

$$\begin{aligned} \hat{Y} &= \hat{\alpha}\hat{\beta}' \\ &= (\frac{1}{m} YZ)(\frac{1}{m^2} Z'Y'YZ)^{-1}(\frac{1}{m} Z'Y') \\ &\quad Y(\frac{1}{n} Y'X)(\frac{1}{n^2} X'Y'YX)^{-1}(\frac{1}{n} X'Y). \end{aligned} \tag{2.19}$$

In practice we can easily find more instruments than the number K of factors, e.g. by the averaging method discussed in Section 2.3 iii). Let us denote by X^* and Z^* extended sets of instruments in number K^*, say, with $K^* \geq K$, such that matrices $E(z_j^* z_j^{*'})$ and $E(x_i^* x_i^{*'})$ are invertible. The model is now overidentified. As usual, formula (2.19) can be applied by selecting K linear combinations of the basic instruments, that is, by considering:

$$X(A) = X^* A, \quad Z(B) = Z^* B, \tag{2.20}$$

where A and B are (K^*, K) full-rank matrices. We get fitted values depending on the selected matrices A and B:

$$\begin{aligned}\hat{Y}(A,B) &= (\frac{1}{m}YZ^*B)(\frac{1}{m^2}B'Z^{*'}Y'YZ^*B)^{-1}(\frac{1}{m}B'Z^{*'}Y') \\ &\quad Y(\frac{1}{n}Y'X^*A)(\frac{1}{n^2}A'X^{*'}Y'YX^*A)^{-1}(\frac{1}{n}A'X^{*'}Y).\end{aligned}$$

How can we choose matrices A and B to make the estimator of $\alpha\beta'$ as accurate as possible? The following Proposition is a direct consequence of the expression of the asymptotic variance given in Proposition 4 and of the standard optimality of the Two Stage Least Squares (2SLS) estimator [see e.g. Gourieroux, Monfort (1995), Property 9.6].

Proposition 6: i) *We have:*

$$\begin{aligned}\min_{A,B} V_{as}\{vec[\sqrt{n}(\hat{Y}(A,B)-\alpha\beta')]\} &= \frac{\sigma^2}{\mu}[\beta\{E(\beta_j z_j^{*'})E(z_j^* z_j^{*'})^{-1}E(z_j^*\beta_j')\}^{-1}\beta'] \otimes Id_n \\ &\quad +\sigma^2 Id_m \otimes [\alpha\{E(\alpha_i x_i^{*'})E(x_i^* x_i^{*'})^{-1}E(x_i^*\alpha_i')\}^{-1}\alpha'],\end{aligned}$$

where the minimization is w.r.t the standard ordering on symmetric matrices.

ii) *This lower bound can be reached in three steps as follows*:

Step 1: *Estimate consistently α and β by applying a double IV method with K instruments selected from the Z^* and X^*.*

Step 2: *Then, estimate the optimal A matrix by a SUR regression of $\hat{\alpha}_i$ on x_i^*, and estimate the optimal B matrix by a SUR regression of $\hat{\beta}_i$ on z_i^*, where $\hat{\alpha}_i$ and $\hat{\beta}_i$ are the first step IV estimators:*

$$\hat{A} = (X^{*\prime}X^*)^{-1}X^{*\prime}\hat{\alpha}, \quad \hat{B} = (Z^{*\prime}Z^*)^{-1}Z^{*\prime}\hat{\beta}.$$

Step 3: *Deduce the generalized double IV estimators of α and β by applying the double 2SLS approach, with instruments $X^*\hat{A}$ and $Z^*\hat{B}$, where $\hat{A}$ and $\hat{B}$ are the estimated optimal selection matrices derived in Step 2.*

Let $\hat{Y}^*$ denote the efficient double IV estimator of the fitted values using the optimal combinations of instruments X^* and Z^*. The asymptotic variance of estimator $\hat{Y}^*$ depends on instruments X^* and Z^* by means of matrices $\{E(\beta_j z_j^{*'})E(z_j^* z_j^{*'})^{-1}E(z_j^*\beta_j')\}^{-1}$ and $\{E(\alpha_i x_i^{*'})E(x_i^* x_i^{*'})^{-1}E(x_i^*\alpha_i')\}^{-1}$,

that are the inverses of the second-order moment matrices of the projection of β_j on z_j^*, and of the projection of α_i on x_i^*, respectively. The minimal asymptotic variance of $\hat{Y}^*$ over all possible sets of instruments can be obtained by an exactly identified set of instruments x_i^* and z_j^* corresponding to factors α_i and β_j, respectively.

Corollary 5: *An optimal choice of the instruments is $x_i^* = \alpha_i$ and $z_j^* = \beta_j$. The corresponding best asymptotic variance of the estimator of the fitted values is:*

$$\min_{X^*,Z^*} V_{as}\{vec[\sqrt{n}(\hat{Y}^* - \alpha\beta')]\} = \frac{\sigma^2}{\mu}[\beta E(\beta_j\beta_j')^{-1}\beta'] \otimes Id_n + \sigma^2 Id_m \otimes [\alpha E(\alpha_i\alpha_i')^{-1}\alpha'].$$

The optimal instruments α_i and β_j are unobservable, but, as in Proposition 6, the efficient estimator of the fitted values corresponding to the optimal choice of the instruments can be computed in two steps. In the first step, we estimate consistently α and β by applying the double IV method based on a set of valid instruments. In the second step, we get the double IV estimators of α and β by applying the double IV approach with instruments $\hat{\alpha}_i$ and $\hat{\beta}_j$ obtained in the first step.

2.7 Comparison with the literature

The literature on large dimensional factor analysis usually estimates the underlying factors by Principal Component Analysis (PCA) [see e.g. Anderson, Rubin (1956), Lawley, Maxwell (1971), Anderson (1984), Stock, Watson (2002), Bai, Ng (2002)]. Typically the PCA estimators of α and β minimize $Tr[(Y - \alpha\beta')'(Y - \alpha\beta')]$ under the normalization restrictions $\frac{1}{m}\beta'\beta = Id_K$ and $\alpha'\alpha$ diagonal, or other types of identifiability restriction asking for instance for different entries of the diagonal matrix $\alpha'\alpha$ ranked in decreasing order [see e.g. Algina (1980), Bekker (1986), Bai, Ng (2013)]. The principal component approach is especially relevant under some normality assumptions, since it corresponds to the maximum likelihood approach [see e.g. Lawley, Maxwell (1971)]. We briefly review Principal Component Analysis in Appendix 3.

It is useful to compare the double IV approach with PCA. For PCA, we have to derive the decreasing sequence of eigenvalues and the associated eigenvectors involved in the singular value decomposition (SVD) of matrix Y. Specifically, the columns of the estimate of the factor matrix β in PCA are the eigenvectors associated with the K largest eigenvalues of matrix $Y'Y$, up to a

normalization (see Appendix 3.1). Equivalently, the PCA estimate of factor matrix α is obtained from the eigenvectors associated with the K largest eigenvalues of matrix YY'. [9]

The standard approach based on PCA has two drawbacks when estimating the vector spaces spanned by the row and column factors, respectively.

i) The first drawback is its computational complexity. For a dataset of large dimensions n, m, with $m = \mu n + o(n)$, $\mu \geq 1$, the number of computations is $O(n^3)$, i.e., it grows cubically in dimension n. This can become a problem for very large data dimensions. Instead, the double IV approach has a numerical complexity of quadratic order $O(n^2)$, i.e. proportional to the number of observations (see Section 2.2).

ii) The second drawback concerns the updating of estimates. The double IV estimators are easily updated with each new datum, that is, when the (n, n) observation matrix becomes a $(n + 1, n + 1)$ observation matrix; in particular they are appropriate for online implementation. Such simple updating does not exist with PCA.

Let us now compare the asymptotic efficiency of the double IV estimator with that of the PCA estimator in terms of fitted values. Let us denote $\hat{Y}^{PCA}$ the estimator of the fitted values matrix $\alpha\beta'$ obtained from the PCA factor estimates. The next Proposition 7 is proved in Appendix 3.2. The proof builds on the asymptotic analysis in e.g. Bai, Ng (2002) and Stock, Watson (2002), but Proposition 7 is novel because it concerns the estimator of the fitted values, instead of the factor estimates themselves.

Proposition 7: *Let $n, m \to \infty$ such that $m = \mu n + o(n)$, $\mu \geq 1$. The PCA estimator of the fitted values $\hat{Y}^{PCA}$ is asymptotically equivalent to the (unfeasible) double IV estimator based on the optimal instruments $x_i = \alpha_i$ and $z_j = \beta_j$.*

From Corollary 5 and Proposition 7, the double IV estimator with optimal instruments is asymptotically equivalent to the PCA estimator for the fitted values. Thus, the double IV approach achieves the same asymptotic efficiency as PCA, but with a reduced degree of computational complexity.

[9] The nonzero eigenvalues of matrices $Y'Y$ and YY' coincide.

3 Dynamic Factor Analysis

Since networks have been around for a long time, there is an abundance of data and we should be expecting more than just a static analysis. For instance, it is interesting to see how the interconnections change over time (and also the clusters, if clusters are introduced, see Section 2.5.2).

3.1 The dynamic factor model

The static factor model (2.1)-(2.2) can be easily extended to a dynamic framework. The observations are matrices of interactions : $Y_t = (y_{i,j,t})$, $t = 1, \ldots, T$, with dimensions (n, m). We assume a factor decomposition with time varying factors and error terms :

$$y_{i,j,t} = \alpha'_{i,t}\beta_{j,t} + \varepsilon_{i,j,t}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m, \quad t = 1, \ldots, T, \tag{3.1}$$

or with matrix notation:

$$Y_t = \alpha_t \beta'_t + \varepsilon_t, \quad t = 1, \ldots, T. \tag{3.2}$$

We assume a fixed number K of factors and make the following assumption, which extends Assumptions A.1-A.2.

Assumption A*.1:

i) *The time series* $(\alpha_{i,t})$, $i = 1, \ldots, n$, $(\beta_{j,t})$, $j = 1, \ldots, m$, *and* $(\varepsilon_{i,j,t})$, $i = 1, \ldots, n$, $j = 1, \ldots, m$ *are independent.*

ii) *These time series are strongly stationary.*

iii) *The time series* $(\alpha_{i,t}), i = 1, \ldots, n$ *[resp* $(\beta_{j,t}), j = 1, \ldots, m$*;* $(\varepsilon_{i,j,t}), i = 1, \ldots, n, j = 1, \ldots, m$*] have identical distributions, such that:*

$$E(\alpha_{i,t}|\underline{\alpha_{i,t-1}}) = 0 \quad [\text{resp. } E(\beta_{j,t}|\underline{\beta_{j,t-1}}) = 0,\ E(\varepsilon_{i,j,t}|\underline{\varepsilon_{i,j,t-1}}) = 0].$$

We assume a big data framework, where the three dimensions n, m and T are large. So the asymptotics is such that $n \to \infty$, $m \to \infty$, and $T \to \infty$.

3.2 Cross-sectional analysis

The static factor analysis of Section 2 can be applied cross-sectionally to get consistent approximations of $\alpha_{i,t}$, $\beta_{j,t}$ and $\varepsilon_{i,j,t}$, $\forall i, j, t$. Let us extend the methodology of asymptotic instruments of Section 2.3 iii) to time series and introduce a transformation of the observation path: $c(y_{i,j,t}, y_{i,j,t-1})$, say, if we just keep the current and lagged observed values. [10] Asymptotic row instruments are:

$$c_{i,.,t} = \frac{1}{m}\sum_{j=1}^{m} c(y_{i,j,t}, y_{i,j,t-1}) \equiv \hat{x}_{i,t}, \tag{3.3}$$

and asymptotic column instruments are:

$$c_{.,j,t} = \frac{1}{n}\sum_{i=1}^{n} c(y_{i,j,t}, y_{i,j,t-1}) \equiv \hat{z}_{j,t}. \tag{3.4}$$

They depend on date t and their values can be gathered in matrices $\hat{X}_t$ and $\hat{Z}_t$. Then, from Proposition 1 we deduce consistent cross-sectional estimators of the factors and errors:

$$\hat{\alpha}_t = \frac{1}{m} Y_t \hat{Z}_t, \quad \hat{\beta}_t = \frac{1}{n} Y_t' \hat{X}_t \hat{C}_t', \quad \hat{\varepsilon}_t = Y_t - \hat{\alpha}_t \hat{\beta}_t', \tag{3.5}$$

with:

$$\hat{C}_t = (\frac{1}{m^2}\hat{Z}_t' Y_t' Y_t \hat{Z}_t)^{-1} (\frac{1}{m}\hat{Z}_t' Y_t') Y_t (\frac{1}{n} Y_t' \hat{X}_t)(\frac{1}{n^2}\hat{X}_t' Y_t Y_t' \hat{X}_t)^{-1}.$$

If dimensions n and m grow at the same rate, these estimators tend to their limits α_t, β_t, ε_t, c_t at rate $1/\sqrt{n}$.

The double IV approach has the important property to select time coherent identification restrictions on the factors, such as:

$$E[z_{j,t}\beta_{j,t}'] = Id_K,$$

at all dates (in the exactly identified case). This time coherency is another advantage of the double IV approach compared to the standard use of PCA. Indeed, the PCA applied date by date does not provide factors with time coherent interpretation. In the double IV approach, the time coherency is imposed by means of the choice of (stationary) instruments.

[10] By considering instruments which are functions of the observable path, we get much more instruments than in the static framework in Section 2.

3.3 Analysis of the factor dynamics

The dynamic factor model (3.1)-(3.2) can be completed by specifying the dynamics of the factors and the errors. This dynamics has to be sufficiently flexible to be compatible with the identification restriction:

$$E(z_{j,t}\beta'_{j,t}) = E[c(y_{i,j,t}, y_{i,j,t-1})\beta'_{j,t}] = Id_K. \tag{3.6}$$

Let us consider the following parametric specifications:

Assumption A*.2: *The series $(\alpha_{i,t})$ [resp. $(\beta_{j,t})$, $(\varepsilon_{i,j,t})$] is a Markov process with continuous transition density $g_\alpha(.|.;\theta_\alpha)$ [resp. $g_\beta(.|.;\theta_\beta)$, $g_\varepsilon(.|.;\theta_\varepsilon)$], where θ_α, θ_β, θ_ε are unknown parameters.*

We have the following Proposition, which is a consequence of general results on Granularity Theory [see e.g. Gagliardini, Gourieroux (2014a, b)].

Proposition 8: *Under Assumptions $A^*.1$-$A^*.2$, the estimators:*

$$\begin{aligned}
\hat{\theta}_\alpha &= \arg\max_{\theta_\alpha} \sum_{i=n}^{n}\sum_{t=1}^{T} \log g_\alpha(\hat{\alpha}_{i,t}|\hat{\alpha}_{i,t-1};\theta_\alpha), \\
\hat{\theta}_\beta &= \arg\max_{\theta_\beta} \sum_{j=1}^{m}\sum_{t=1}^{T} \log g_\beta(\hat{\beta}_{j,t}|\hat{\beta}_{j,t-1};\theta_\beta), \\
\hat{\theta}_\varepsilon &= \arg\max_{\theta_\varepsilon} \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{t=1}^{T} \log g_\varepsilon(\hat{\varepsilon}_{i,j,t}|\hat{\varepsilon}_{i,j,t-1};\theta_\varepsilon),
\end{aligned}$$

are consistent, asymptotically normal and asymptotically efficient.

As in the static case, adjustments in the objective criteria are needed if the (joint) distribution allows for point mass at zero for the values of date t and date $t-1$.

3.4 Nested factor models

Model (3.1) includes factors α (resp. β) doubly indexed by (i,t) [resp (j,t)]. This model can be constrained to get factors indexed by either i, or j, or t, only. Such a constrained model can be defined as :

$$y_{i,j,t} = \sum_{k=1}^{K} \alpha_{i,t,k}\beta_{j,t,k} + \varepsilon_{i,j,t}, \tag{3.7}$$

where:

$$\begin{aligned}\alpha_{i,t,k} &\equiv \sum_{l=1}^{L(\alpha_k)} a_{i,l}(\alpha_k) b_{t,l}(\alpha_k), \\ \beta_{j,t,k} &\equiv \sum_{l=1}^{L(\beta_k)} a_{j,l}(\beta_k) b_{t,l}(\beta_k). \end{aligned} \tag{3.8}$$

We deduce the observations as functions of the new factors a and b:

$$y_{i,j,t} = \sum_{k=1}^{K} \sum_{l=1}^{L(\alpha_k)} \sum_{l^*=1}^{L(\beta_k)} [a_{i,l}(\alpha_k) a_{j,l^*}(\beta_k) b_{t,l}(\alpha_k) b_{t,l^*}(\beta_k)] + \varepsilon_{i,j,t}, \tag{3.9}$$

with $\sum_{k=1}^{K} L(\alpha_k)$ factors indexed by i, $\sum_{k=1}^{K} L(\beta_k)$ factors indexed by j.

The double IV estimation approach provides consistent estimators of this constrained model including more factors indexed by $i, j,$ or t. They are obtained in two steps :

Step 1: Apply the double IV approach to the observations $Y_t = (y_{i,j,t})$ and deduce the estimated $\hat{\alpha}(k) = (\hat{\alpha}_{i,t,k}), \hat{\beta}(k) = (\hat{\beta}_{j,t,k}), k = 1, \ldots, K.$

Step 2: Apply the double IV approach to the pseudo-observations $\hat{\alpha}(k), k = 1, \ldots, K$ [resp. $\hat{\beta}(k), k = 1, \ldots, K$] to deduce the $\hat{a}_{i,l}(\alpha_k), \hat{b}_{t,l}(\alpha_k)$ [resp. $\hat{a}_{j,l}(\beta_k), \hat{b}_{t,l}(\beta_k)$].

4 Experimental results

To illustrate the double IV approach, we apply the methodology to artificial data sets.

4.1 One-factor model with mixtures of discrete-continuous distributions

We first consider a single static factor model:

$$y_{i,j} = \alpha_i \beta_j + \varepsilon_{i,j}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m, \tag{4.1}$$

where the factors $\alpha_i' s$, $\beta_j' s$ and errors $\varepsilon_{i,j}' s$ are drawn independently in mixtures of point masses at zero and log-normal distributions along the two schemes in Table 1. The schemes differ in terms of the parameters of the log-normal distribution of the errors, with larger and more volatile errors in scheme 2.

Table 1: Simulation schemes 1 and 2.

	weight on point mass at 0	parameters of the log-normal distribution	
		μ	σ^2
α	0.4	0	1
β	0.4	1	2
ε, scheme 1	0.4	1	2
ε, scheme 2	0.4	3	2

The schemes in Table 1 are both compatible with zero values for the observable variable y, with probability:

$$\begin{aligned} P[y_{i,j}=0] &= P[\varepsilon_{i,j}=0]P[\text{either } \alpha_i=0, \text{ or } \beta_j=0] \\ &= P[\varepsilon_{ij}=0]\{1-P[\alpha_i\neq 0]P[\beta_j\neq 0]\} \\ &= 0.4(1-0.6^2)=0.256. \end{aligned} \tag{4.2}$$

We first draw independently the $\alpha's$, $\beta's$ and $\varepsilon's$ along scheme 1 in Table 1 with dimensions $n=m=10,000$. In such a big data environment, it is difficult to represent the complete set of available data. However, summary statistics can be informative. We provide in the upper panel of Figure 1 the North-West submatrix of Y with size (100,100) to highlight the effect of zero values. Each observation is represented by a dot, whose color depends on the magnitude $y_{i,j}$.

[Insert Figure 1: Submatrices of observations (schemes 1 and 2)]

We observe vertical and horizontal patterns of large observations, corresponding to individuals with large β and α factors, respectively. The vertical patterns are more pronounced because of the larger scale of the β factor.

Other summary statistics use more observations. We provide in Figures 2, 3 and 4 the sample distributions of $y_{i,1}$ and $y_{i,2}$, for $i=1,\ldots,10000$, the sample distributions of $y_{1,j}$ and $y_{2,j}$, for $j=1,\ldots,10000$, and the sample distribution of $y_{i,j}$, for $i,j=1,\ldots,10000$, respectively.

[Insert Figure 2: Sample distributions of $y_{i,1}$ and $y_{i,2}$ (scheme 1)]

[Insert Figure 3: Sample distributions of $y_{1,j}$ and $y_{2,j}$ (scheme 1)]

[Insert Figure 4: Sample distribution of $y_{i,j}$ (scheme 1)]

It is interesting to compare the sample distributions of $y_{i,1}$ and $y_{i,2}$, say. Indeed, they approximate the distributions of $y_{i,j}$ given β_j, for $j = 1, 2$. These distributions differ since $\beta_1 \neq \beta_2$. More precisely, we have $\beta_1 = 2.89$ and $\beta_2 = 0$ in the simulated dataset, which explains the larger scale of observations $y_{i,1}$ compared to observations $y_{i,2}$ in Figure 2. Moreover, since $\alpha_2 = \beta_2 = 0$, the distributions of $y_{i,2}$ and $y_{2,j}$ are equal, and correspond to the mixture distribution of $\varepsilon_{i,j}$. Let us now consider the distributions of $y_{i,1}$ and $y_{1,j}$. Since $\beta_1 \neq 0$ and $\alpha_1 \neq 0$, we have $P[y_{i,1} = 0] = P[\alpha_i = 0]P[\varepsilon_{i,j} = 0] = 0.16 = P[y_{1,j} = 0]$, and the continuous parts of these distributions correspond to sums of log-normal variables. Finally, in Figure 4 the proportion of observations $y_{i,j} = 0$ is close to 0.25 as implied by (4.2).

We apply the double IV approach to this artificial data set, and compute the estimates $\hat{\alpha}_i$, $\hat{\beta}_j$, $\hat{\varepsilon}_{i,j}$. In the single-factor model we need a single row and a single column instrument. We choose $x_i = y_{i,\cdot}$ and $z_j = y_{\cdot,j}$ corresponding to the identity transform $a(.)$ in Section 2.3 iii). Since factors and errors do not have zero means, we apply the ANOVA transformation to the y data [see Section 2.3 i)]. The computation of the estimates for all sample units requires about 2 seconds on a standard computer. The scatter plots in the three upper panels of Figure 5 show that both the factor estimates $\hat{\alpha}_i$, $\hat{\beta}_j$ and the fitted values $\hat{y}_{i,j} = \hat{\alpha}_i\hat{\beta}_j$ are close to the corresponding true values, uniformly across the sample. This remark is confirmed by the comparison of the North-West (100,100) submatrices of fitted values $\hat{\alpha}\hat{\beta}'$ and true values $\alpha\beta'$, displayed in the upper left and right panel of Figure 6, respectively.

[Insert Figure 5: Scatter plots of estimates vs true values (schemes 1 and 2)]

[Insert Figure 6: Submatrices of fitted values (schemes 1 and 2)]

To assess the effect of the size of errors on the estimators accuracy, we simulate a second artificial data set along scheme 2 in Table 1, with dimensions $n = m = 10000$, and apply the double IV approach with the same instruments as above. The North-West submatrix of Y with size (100,100) is displayed in the lower panel of Figure 1, and the scatter plots of estimates versus true

values are displayed in the three lower panels of Figure 5. The North-West (100,100) submatrix of fitted values is displayed in the lower left panel of Figure 6. Comparing the two panels of Figure 1, under simulation scheme 2 the vertical and horizontal patterns induced by the factor structure are less visible due to the larger size of the errors. The estimators are less accurate on the simulated dataset under scheme 2 than under scheme 1. In the three lower panels of Figure 5, the estimators tend to be more accurate for large values of α_i and β_j, i.e., when the impact of errors is smaller in relative terms.

4.2 One-factor model with missing observations

Let us now consider a one-factor model with missing data. The observations are:

$$y_{i,j} = (\alpha_i \beta_j + \varepsilon_{i,j})\xi_{i,j}, \quad i = 1, ..., n,\ j = 1, ..., m.$$

The factors α, β and the errors ε are drawn from log-normal distributions with parameters given in Table 2. We consider two simulation schemes for the indicator variable ξ, corresponding to rates of missing observations equal to 10% and 40% in schemes 3 and 4, respectively.

Table 2: Simulation schemes 3 and 4.

	parameters of the log-normal distribution	
	μ	σ^2
α	0	1
β	1	2
ε	1	2
	probability of missing data	
scheme 3	$P(\xi_{i,j} = 0) = 0.90$	
scheme 4	$P(\xi_{i,j} = 0) = 0.60$	

The sample sizes are $n = m = 10000$ as in Section 4.1. We apply the double IV method after performing the ANOVA transformation in (2.18), with instruments $x_i = y_{i,\cdot}$ and $z_j = y_{\cdot,j}$.

[Insert Figure 7: Submatrices of observations (schemes 3 and 4)]

[Insert Figure 8: Scatter plots of estimates vs true values (schemes 3 and 4)]

[Insert Figure 9: Submatrices of fitted values (schemes 3 and 4)]

As expected, the rate of missing observations is larger in the data matrix simulated under scheme 4 (Figure 7). This explains the smaller accuracy of the factor estimates under this scheme (Figure 8). Moreover, in both schemes 3 and 4 the estimates of factor α are less accurate than those of factor β.

5 Concluding remarks

Factor models are introduced to reduce the dimension of the analysis. In the static interaction model, we pass from dimension 2, i.e. the dimension of the observations, to dimension 1, i.e. the dimension of the factors. In the dynamic interaction model from dimension 3 to dimension 2.

The aim of our paper was to answer the question [11] : "Are more data always better for factor analysis ?". For huge data sets the standard factor analysis can become numerically complicated. However we can benefit from big data to introduce new techniques, which are much less computational demanding. The double instrumental variable approach is an example of such a technique able to reach the same asymptotic efficiency as Principal Component Analysis. We have also explained how asymptotic instruments can be constructed from the bidimensional endogenous observations.

We have shown that the double IV approach is easily extended to the case of incomplete data. The approach is an alternative to other collaborative filtering methods either based on nuclear penalized estimators, similarities between individuals, or on block structured models with Bayesian estimation.

We have explained how to derive by simple explicit formulas the parameters of interest, including the factor values and their distributions. In a framework of online data, it would be interesting to consider explicitly how to update these estimates at each new data arrival. This learning aspect is left for future research.

[11] appearing as the title of the paper by Boivin, Ng (2006).

REFERENCES

ACM SIGKDD and Netflix, (2007) : Proceedings of KDD Cup and Workshop. Available online at http://www.cs.vic.edu/ livb/KDD-cup-2007/proceedings.html.

Algina, J., (1980) : "A Note on Identification in the Oblique and Orthogonal Factor Analysis Models", Psychometrika, 45, 393-395.

Anderson, T. (1984) : *An Introduction to Multivariate Statistical Analysis*, New-York, Wiley.

Anderson, T., and H., Rubin (1956) : "Statistical Inference in Factor Analysis", in J., Neyman (ed.) : Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol 5., Berkeley, University of California Press, 114-150.

Bai, J., and S., Ng (2002): "Determining the Number of Factors in Approximate Factor Models", Econometrica, 70, 191-221.

Bai, J., and S., Ng (2008) : "Large Dimensional Factor Analysis", in *Foundations and Trends in Econometrics*, Vol 3, 89-163.

Bai, J., and S., Ng (2010) : "Instrumental Variables Estimation in a Data Rich Environment", Econometric Theory, 26, 1607-1637.

Bai, J., and S., Ng (2013) : "Principal Components Estimation and Identification of Static Factors", Journal of Econometrics, 176, 18-29.

Bekker, P. (1986) : "A Note on the Identification of Restricted Factor Loading Matrices", Psychometrika, 51, 607-611.

Belloni, A., Chernozhukov, V., and L., Wang (2011) : "Square-Root Lasso : Pivotal Recovery of Sparse Signals via Conic Programming", Biometrika, 98, 791-806.

Boivin, J., and S., Ng (2006) : "Are More Data Always Better for Factor Analysis ?", Journal of Econometrics, 132, 169-194.

Boucher, V., and I., Mourifie (2013) : ”My Friend Far Far Away : Asymptotic Properties of Pairwise Stable Networks”, DP University of Toronto.

Candes, E., and Y., Plan (2010) : ”Matrix Completion with Noise”, Proceedings of the IEEE, 98, 925-936.

Candes, E., and B., Recht (2009) : ”Exact Matrix Completion via Convex Optimization”, Found. Comput. Math., 9, 717-772.

Chen, H. (2002): “Principal Component Analysis with Missing Data and Outliers”, working paper.

Comon, P. (1994): “Independent Component Analysis: A New Concept?”, Signal Processing, 36, 287-314.

Davies, P. (2012) : ”Interactions in the Analysis of Variance”, JASA, 107, 1502-1509.

Dawid, A. (1981) : ”Some Matrix-Variate Distribution Theory : Notational Considerations and a Bayesian Application”, Biometrika, 68, 265-274.

Forni, M., and L., Reichlin (1996) : ”Dynamic Common Factors in Large Cross-Sections”, Empirical Economics, 21, 27-42.

Frank, O., and D., Strauss (1986) : ”Markov Graphs”, JASA, 81, 832-842.

Gagliardini, P., and C., Gourieroux (2014a): *Granularity Theory with Applications to Finance and Insurance*, Cambridge University Press.

Gagliardini, P., and C., Gourieroux (2014b): “Efficiency in Large Dynamic Panel Models with Common Factors”, Econometric Theory, 30, 961-1020.

Geman, F. (1980): “A Limit Theorem for the Norm of Random Matrices”, Annals of Probability, 8, 252-261.

Goldberg, D., Nichols, D., Oki, B., and D., Terry (1992) : ”Using Collaborative Filtering to

Weave an Information Tapestry", Communications of ACM, 35, 61-70.

Gourieroux, C., Heam, J.C., and A., Monfort (2012) : "Bilateral Exposures and Systemic Solvency Risk", Canadian Journal of Economics, 45, 1273-1309.

Gourieroux, C., and A., Monfort (1995): *Statistics and Econometric Models*, Cambridge University Press.

Gourieroux, C., Monfort, A., and A., Trognon (1982): "Moindre Carrés Asymptotiques", Annales de l'ENSAE, 58, 91-122.

Granger, C. (1987) : "Implication of Aggregation with Common Factors", Econometric Theory, 3, 208-222.

Gross, D. (2011) : "Recovering Low-Rank Matrices From Few Coefficients in Any Basis", IEEE Transactions, 57, 1548-1566.

Gupta, A., and K., Nagar (2000) : *Matrix Variate Distributions*, Boca Raton FL, Chapman and Hall.

Handcock, M., Raftery, A., and J., Tantrum (2007) : "Model-Based Clustering for Social Networks", with discussion, JRSS A, 170, 301-354.

Hoff, P., Raftery, A., and M., Handcock (2002) : "Latent Space Approaches to Social Network Analysis", JASA, 97, 1090-1098.

Hofmann, T. (2001) : "Unsupervised Learning by Probabilistic Latent Semantic Analysis", Machine Learning, 42, 177-196.

Hofmann, T. (2003) : "Collaborative Filtering via Gaussian Probabilistic Latent Semantic Analysis", Proceedings of the ACM-SIGIR.

Hofmann, T. (2004) : "Latent Semantic Models for Collaborative Filtering", ACM Transactions on Information Systems, 22, 89-115.

Holland, P., and S., Leinhardt (1981) : ”An Exponential Family of Probability Distributions for Directed Graphs”, with discussion, JASA, 76, 33-65.

Horn, R., and C., Johnson (1985): *Matrix Analysis*, Cambridge University Press.

Hyvarinen, A., Karhunen, J., and E., Oja (2001): *Independent Component Analysis*, New York, Wiley.

Iijima, R., and Y., Kamada (2010) : ”Social Distance and Network Structures”, DP Harvard University.

Jackson, M. (2008) : *Social and Economic Networks*, Princeton University Press.

Jolliffe, I. (2002): *Principal Component Analysis*, Springer Series in Statistics.

Keshavan, R., Montanari, A., and S., Oh (2010) : ”Matrix Completion From Noisy Entries”, J. Mach. Learn. Res., 11, 2057-2078.

Klopp, O. (2014) : ”Noisy Low Rank Matrix Completion with General Sampling Distribution”, Bernoulli, 2, 282-303.

Kodde, D., Palm, F., and G., Pfann (1990): “Asymptotic Least Squares Estimation: Efficiency Considerations and Applications”, Journal of Applied Econometrics, 5, 229-243.

Koltchinskii, V., Tsybakov, A., and K., Lounici (2011) : ”Nuclear-Norm Penalization and Optimal Rates for Noisy Low-Rank Matrix Completion”, Annals of Statistics, 39, 2302-2329.

Kranton, R., and D., Minehart (2011) : ”A Theory of Buyer-Seller Networks”, The American Economic Review, 91, 485-508.

Latouche, P., Birmele, E., and C., Ambroise (2011) : ”Overlapping Stochastic Block Models with Application to the French Political Blogosphere”, Annals of Applied Statistics, 5, 309-336.

Lawley, D., and A., Maxwell (1971) : *Factor Analysis in a Statistical Method*, Butterworth, London.

Leng, C., and C., Tang (2012) : ”Sparse Matrix Graphical Models”, JASA, 107, 1187-1200.

Magnus, J., and H., Neudecker (1994) : *Matrix Differential Calculus with Applications in Statistics and Econometrics*, 2nd ed., Wiley.

Miyahara, K., and M., Pazzani (2002) : ”Improvement of Collaborative Filtering with the Simple Bayesian Classific”, Information Processing Society of Japan, 43, 11.

Negahban, S., and M., Wainwright (2012) : ”Restricted Strong Convexity and Weighted Matrix Completion : Optimal Bounds with Noise”, J. Mach. Learn. Res., 13, 1665-1697.

Newman, M. (2003) :”The Structure and Function of Complex Networks”, SIAM Rev., 45, 167-256.

Nowicki, K., and T., Snijders (2005) : ”Estimation and Prediction for Stochastic Blockstructures”, JASA, 96, 1077-1087.

Recht, B. (2011) : ”A Simpler Approach to Matrix Completion”, Journal of Machine Learning Research, 12, 3413-3430.

Stock, J., and M., Watson (2002) : ”Forecasting Using Principal Components From a Large Number of Predictors”, JASA, 97, 1167-1179.

Su, X., and T., Khoshgoftaar (2009) : ”A Survey of Collaborative Filtering Techniques”, in *Advances in Artificial Intelligence*, Hindawc Publishing Corporation.

Suhr, D. (2009) : ”Principal Component Analysis vs Exploratory Factor Analysis”, SUGI 30 Proceedings.

Theil, H. (1953): *Repeated Least Squares Applied to Complete Equations Systems*, The Hague, Central Planning Bureau.

Traxillo, C. (2003) : ”Multivariate Statistical Methods : Practical Research Applications Courses Notes”, Cary, N.C. : SAS Institute.

Upper, C., and A., Worms (2004) : ”Estimating Bilateral Exposures in the German Interbank Market : Is There a Danger of Contagion ?”, European Economic Review, 48, 827-849.

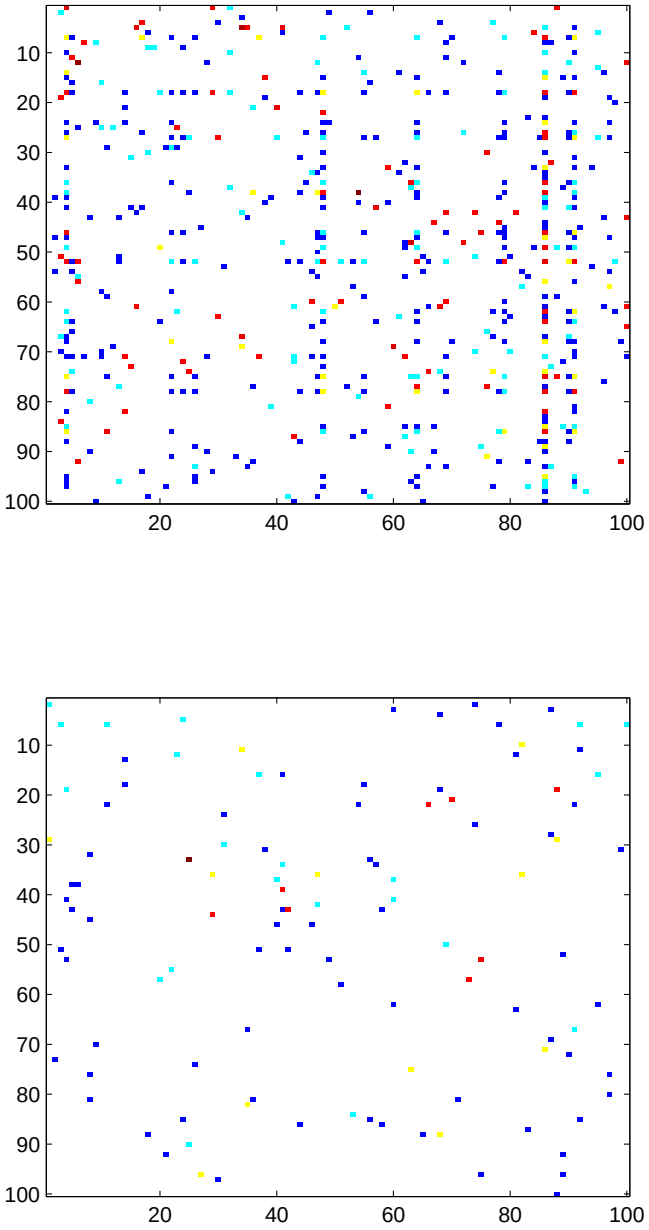
Wang, P. (2008) : ”Large Dimensional Factor Models with a Multi-Level Factor Structure : Identification, Estimation and Inference”, New-York University DP.

Wasserman, S., and C., Anderson (1987) : ”Stochastic a Posteriori Blockmodels: Construction and Assessment”, Social Networks, 9, 1-36.

Wasserman, S., and K., Faust (1994) : ”Social Network Analysis : Methods and Applications”, Cambridge University Press.

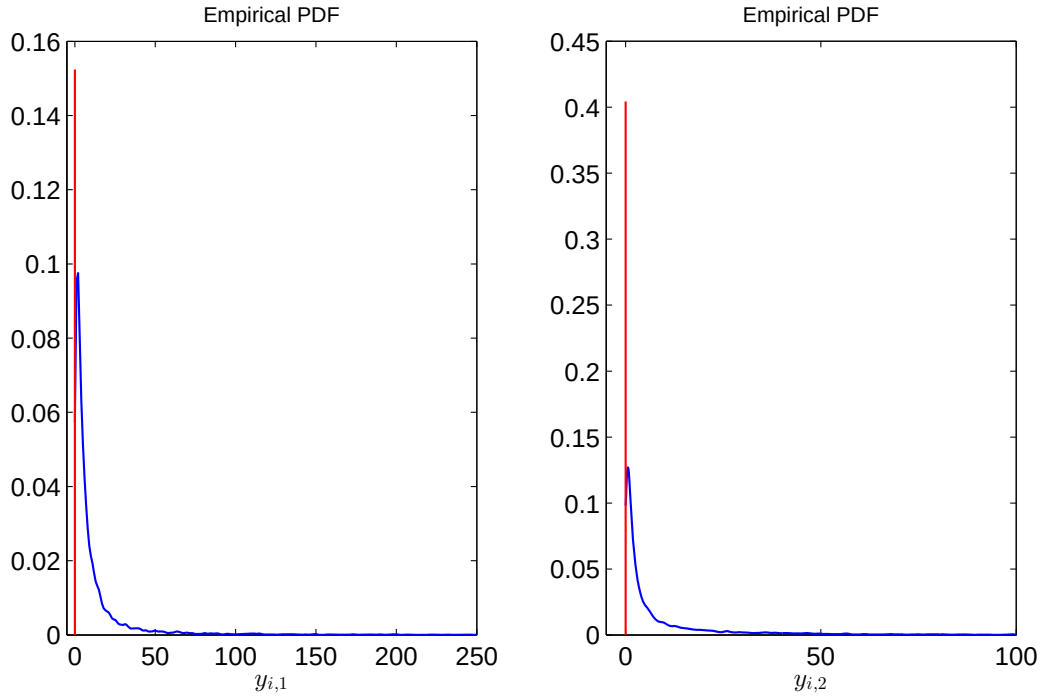
Yin, Y., Bai, Z., and P., Krishnaiah (1988): “On the Limit of the Largest Eigenvalue of the Large Dimensional Sample Covariance Matrix”, Probability Theory, 78, 509-521.

Figure 1: Submatrices of observations (schemes 1 and 2).



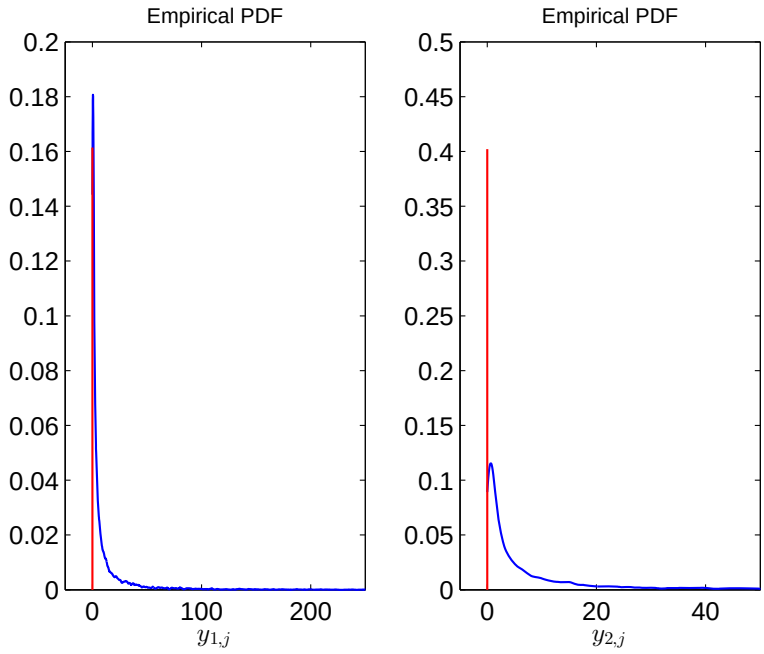
Each panel of this Figure displays the North-West $(100, 100)$ block of a simulated data matrix $Y = (y_{i,j})$ in the static one-factor model of Section 4.1. The distributions of factors α_i, β_j and errors $\varepsilon_{i,j}$ are mixtures of point masses at zero and log-normal distributions with parameters given in Table 1, for scheme 1 in the upper panel, and for scheme 2 in the lower panel, respectively. For the cell on row i and column j, the color of the dot is related to the magnitude of observation $y_{i,j}$, such that cold (resp. hot) colors correspond to small (resp. large) observations.

Figure 2: Sample distributions of $y_{i,1}$ and $y_{i,2}$ (scheme 1).



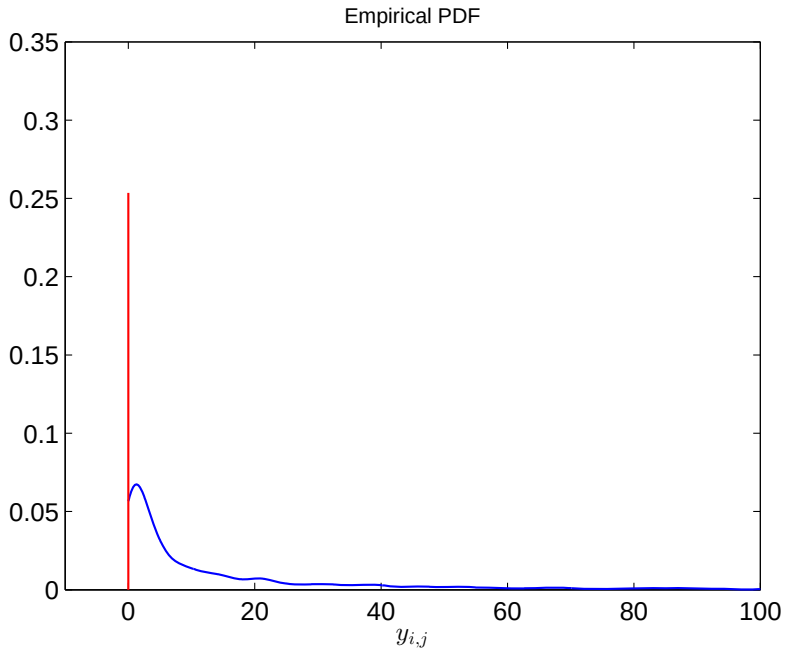
The left and right panels display estimates of the mixture probability distribution functions of variables $y_{i,1}$ and $y_{i,2}$, respectively. In each panel, the height of the vertical bar at zero is equal to the sample proportion of zero values. The blue curve is a kernel density estimate computed on the subsample of nonzero values, multiplied by the proportion of nonzero values. The data are generated according to scheme 1 in Table 1.

Figure 3: Sample distributions of $y_{1,j}$ and $y_{2,j}$ (scheme 1).



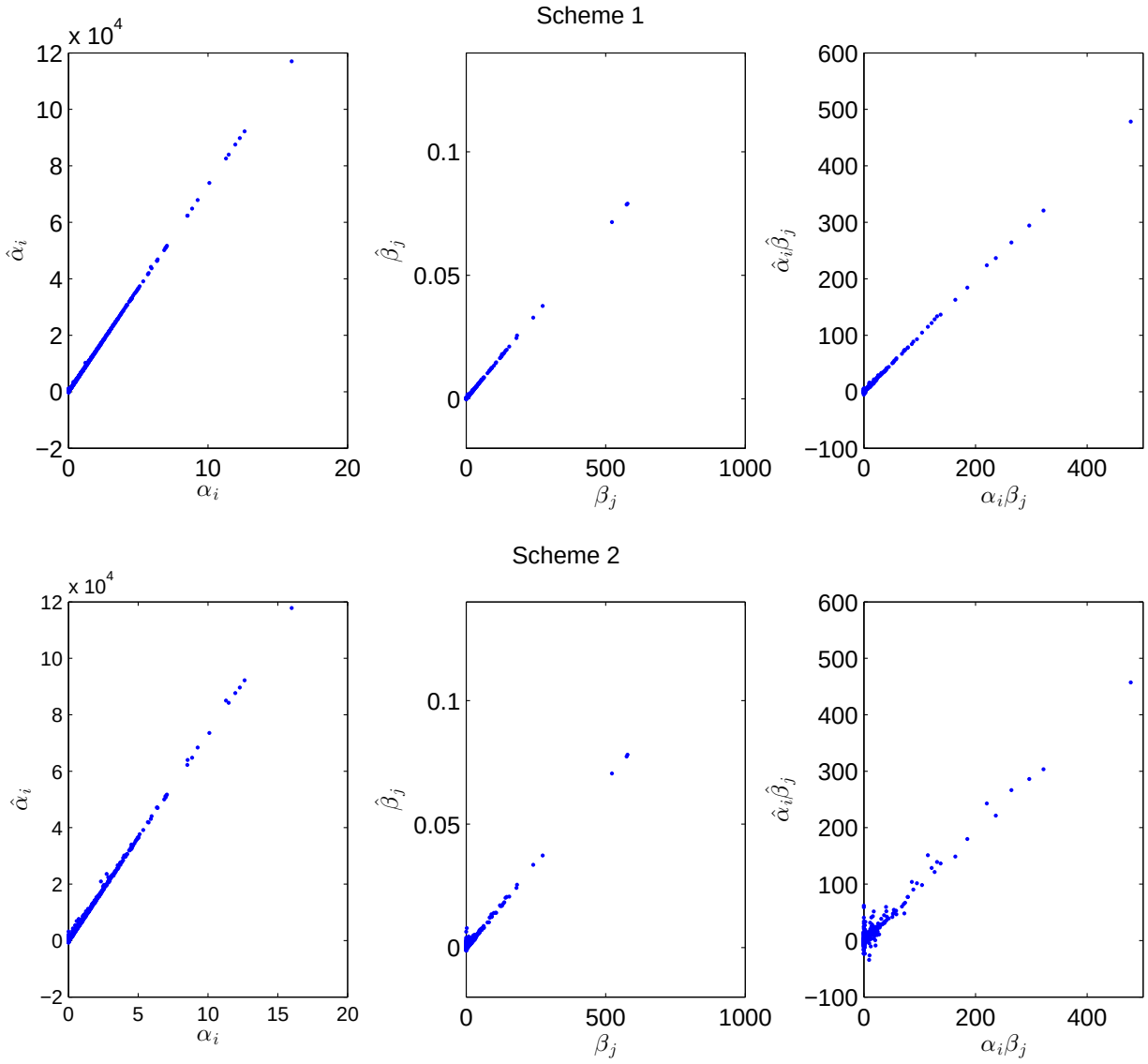
The left and right panels display estimates of the mixture probability distribution functions of variables $y_{1,j}$ and $y_{2,j}$, respectively. In each panel, the height of the vertical bar at zero is equal to the sample proportion of zero values. The blue curve is a kernel density estimate computed on the subsample of nonzero values, multiplied by the proportion of nonzero values. The data are generated according to scheme 1 in Table 1.

Figure 4: Sample distribution of $y_{i,j}$ (scheme 1).



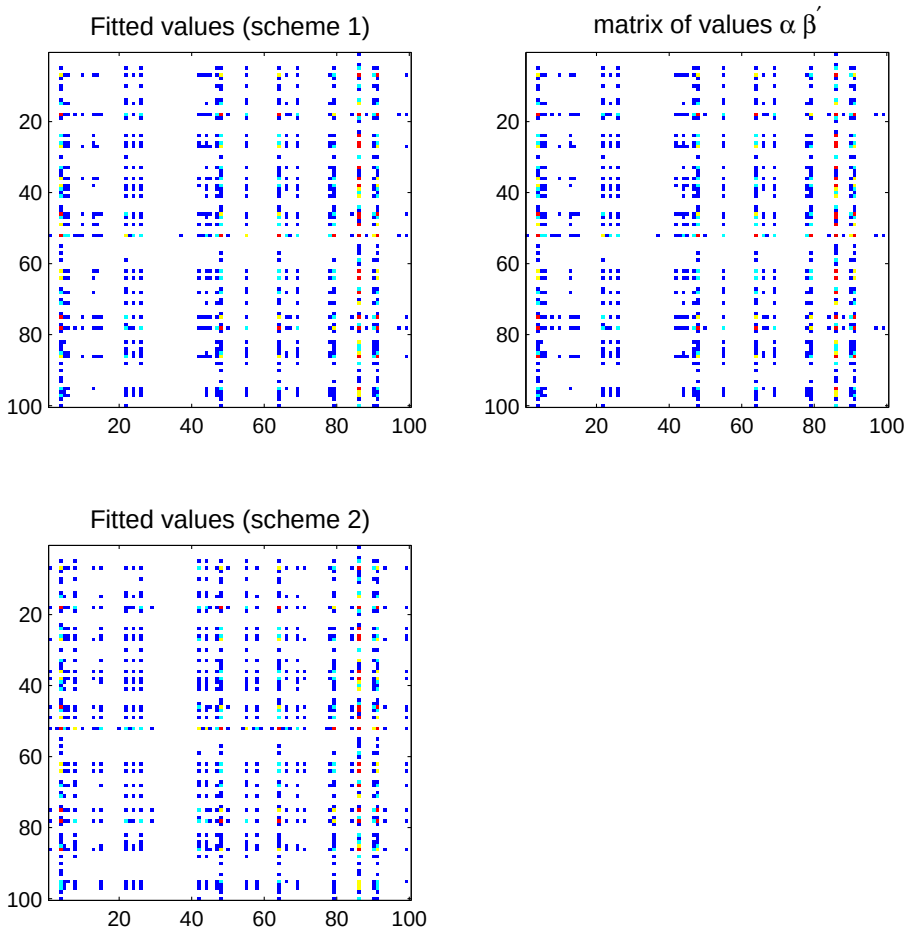
The panel displays the estimate of the mixture probability distribution function of variables $y_{i,j}$. The height of the vertical bar at zero is equal to the sample proportion of zero values. The blue curve is a kernel density estimate computed on the subsample of nonzero values, multiplied by the proportion of nonzero values. The data are generated according to scheme 1 in Table 1.

Figure 5: Scatter plots of estimates vs true values (schemes 1 and 2).



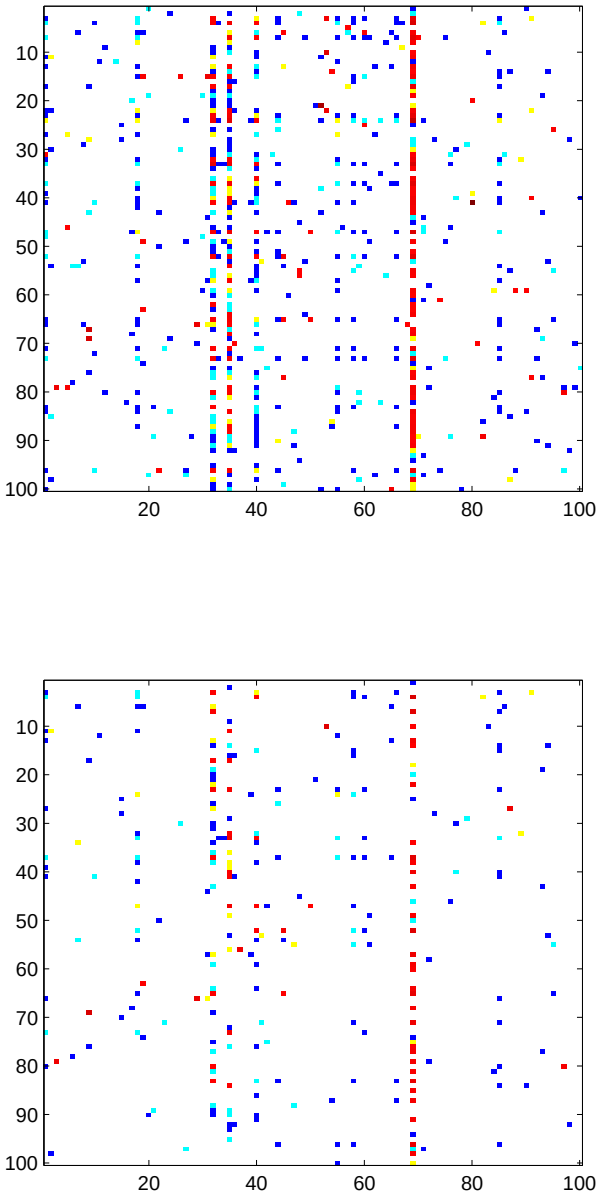
The left, middle and right panels display scatter plots of estimates vs true values for $(\alpha_i, \hat{\alpha}_i)$, $(\beta_j, \hat{\beta}_j)$, and $(\alpha_i\beta_j, \hat{\alpha}_i\hat{\beta}_j)$, respectively. The estimates are obtained with the double IV method using instruments $x_i = y_{i,\cdot}$ and $z_j = y_{\cdot,j}$. The data are generated according to scheme 1 in Table 1 for the three upper panels, and according to scheme 2 in Table 1 for the three lower panels.

Figure 6: Submatrices of fitted values (schemes 1 and 2).



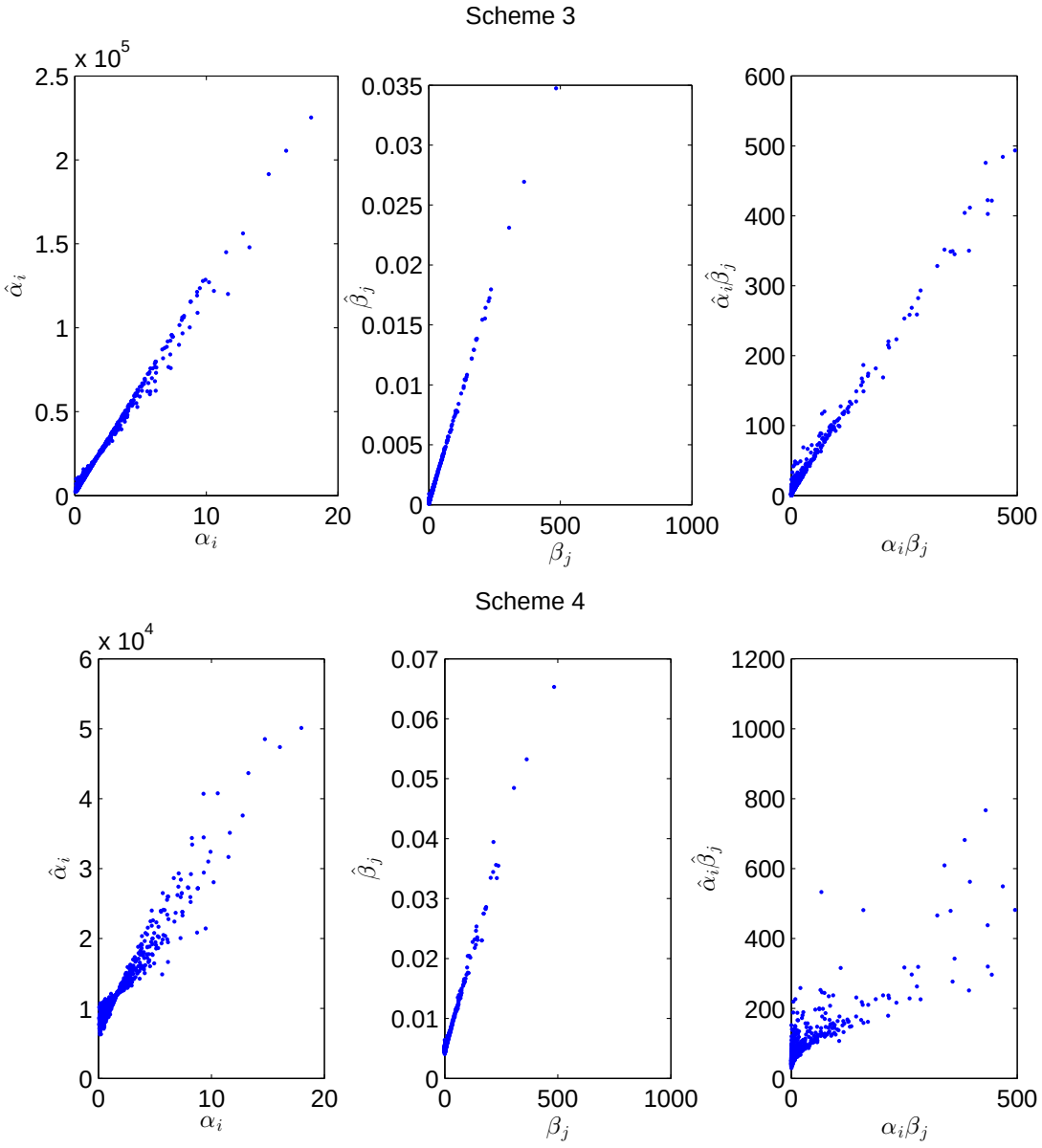
The upper left panel of this Figure displays the North-West $(100, 100)$ block of the matrix of fitted values $\hat{y}_{i,j} = \hat{\alpha}_i \hat{\beta}_j$ in the static one-factor model of Section 4.1. The distributions of factors α_i, β_j and errors $\varepsilon_{i,j}$ are mixtures of point masses at zero and log-normal distributions with parameters given in Table 1, scheme 1. The estimates are obtained with the double IV method using instruments $x_i = y_{i,\cdot}$ and $z_j = y_{\cdot,j}$. The lower left panel displays the North-West $(100, 100)$ block of the matrix of fitted values $\hat{y}_{i,j} = \hat{\alpha}_i \hat{\beta}_j$ for data generated with parameters given in Table 1, scheme 2. The upper right panel displays the North-West $(100, 100)$ block of matrix $\alpha\beta'$. For the cell on row i and column j, the color of the dot is related to the magnitude of fitted value $\hat{y}_{i,j}$, such that cold (resp. hot) colors correspond to small (resp. large) values.

Figure 7: Submatrices of observations (schemes 3 and 4) .



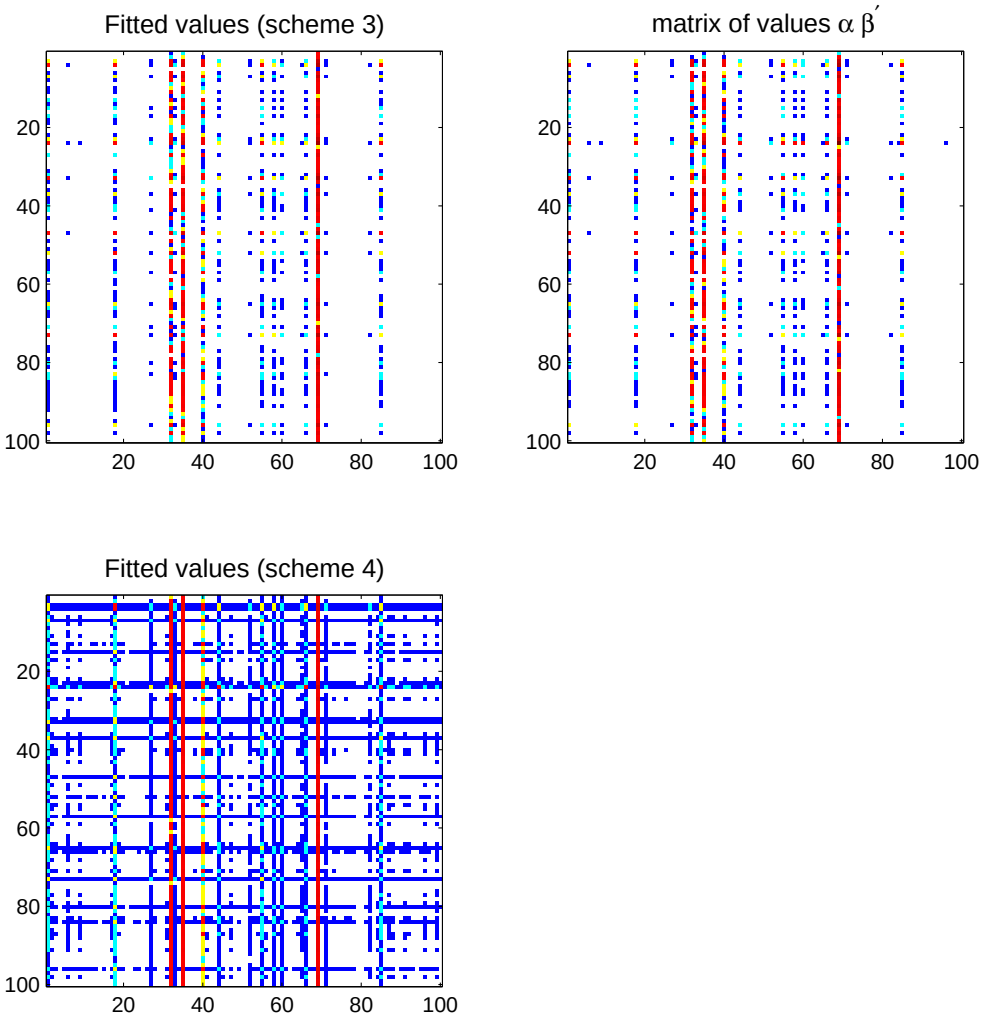
Each panel of this Figure displays the North-West $(100, 100)$ block of a simulated data matrix $Y = (y_{i,j})$ in the static one-factor model with missing data of Section 4.2. The distributions of factors α_i, β_j and errors $\varepsilon_{i,j}$ are log-normal distributions with parameters given in Table 2, and the indicator variables $\xi_{i,j}$ are i.i.d. Bernoulli distributed with parameter 0.90 (scheme 3) in the upper panel, and Bernoulli parameter 0.60 (scheme 4) in the lower panel. For the cell on row i and column j, the color of the dot is related to the magnitude of observation $y_{i,j}$, such that cold (resp. hot) colors correspond to small (resp. large) observations.

Figure 8: Scatter plots of estimates vs true values (schemes 3 and 4).



The left, middle and right panels display scatter plots of estimates vs true values for $(\alpha_i, \hat{\alpha}_i)$, $(\beta_j, \hat{\beta}_j)$, and $(\alpha_i\beta_j, \hat{\alpha}_i\hat{\beta}_j)$, respectively. The estimates are obtained with the double IV method using instruments $x_i = y_{i,\cdot}$ and $z_j = y_{\cdot,j}$. The data are generated according to scheme 3 in Table 2 for the three upper panels, and according to scheme 4 in Table 2 for the three lower panels.

Figure 9: Submatrices of fitted values (schemes 3 and 4).



The upper and lower left panels of this Figure display the North-West $(100, 100)$ block of the matrix of fitted values $\hat{y}_{i,j} = \hat{\alpha}_i \hat{\beta}_j$ in the static one-factor model with missing data of Section 4.2. The distributions of factors α_i, β_j and errors $\varepsilon_{i,j}$ are log-normal distributions with parameters given in Table 2, and the indicator variables $\xi_{i,j}$ are i.i.d. Bernoulli distributed with parameter 0.90 (scheme 3) in the upper panel, and with Bernoulli parameter 0.60 (scheme 4) in the lower panel. The estimates are obtained with the double IV method using instruments $x_i = y_{i,\cdot}$ and $z_j = y_{\cdot,j}$. The upper right panel displays the North-West $(100, 100)$ block of matrix $\alpha\beta'$. For the cell on row i and column j, the color of the dot is related to the magnitude of fitted value $\hat{y}_{i,j}$, such that cold (resp. hot) colors correspond to small (resp. large) values.

Appendix 1: Asymptotic Expansions

A.1.1 Proof of the probability limits in (2.3) and (2.4)

We have:

$$\begin{aligned}\operatorname*{plim}_{n\to\infty}\frac{1}{n}X'Y &= \operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}x_i y_i' \\ &= E(x_i y_i') \\ &= E(x_i\alpha_i'\beta') + E(x_i\varepsilon_i') \\ &= E(x_i\alpha_i')\beta'.\end{aligned}$$

The proof is similar for the limit $\operatorname*{plim}_{m\to\infty}\frac{1}{m}Z'Y' = C^*(z,\beta)\alpha'$.

A.1.2 Matrix expression of the double instrumental variable estimator

We will use the following properties of the Kronecker product [see e.g. Magnus, Neudecker (1994)]:

$$(A\otimes B)(C\otimes D) = (AC)\otimes(BD), \tag{a.1}$$

$$(A\otimes B)' = A'\otimes B', \tag{a.2}$$

$$(A\otimes B)^{-1} = (A^{-1})\otimes(B^{-1}), \tag{a.3}$$

where matrices A and B are non-singular in the last equality. We deduce from Proposition 1 i):

$$\begin{aligned}vec\,\hat{C} &= \left\{\left[(\frac{1}{n}X'Y)\otimes(\frac{1}{m}Z'Y')\right]\left[(\frac{1}{n}Y'X)\otimes(\frac{1}{m}YZ)\right]\right\}^{-1}\left[(\frac{1}{n}X'Y)\otimes(\frac{1}{m}Z'Y')\right]vecY \\ &= \left\{\left[(\frac{1}{n^2}X'YY'X)^{-1}(\frac{1}{n}X'Y)\right]\otimes\left[(\frac{1}{m^2}Z'Y'YZ)^{-1}(\frac{1}{m}Z'Y')\right]\right\}vec\,Y,\end{aligned}$$

by (a.1), (a.2), (a.3). Finally, by applying (2.8), we get:

$$\hat{C} = \left(\frac{1}{m^2}Z'Y'YZ\right)^{-1}\left(\frac{1}{m}Z'Y'\right)Y\left(\frac{1}{n}Y'X\right)\left(\frac{1}{n^2}X'YY'X\right)^{-1}. \tag{a.4}$$

A.1.3 Consistency of the double IV estimator of C

Before deriving the limit of $\hat{C}$, let us recall the identification restriction (2.5):

$$C^*(z,\beta) = E(z_j\beta_j') = Id_K, \tag{a.5}$$

and its consequences for the factor interpretations [see (2.3)-(2.5)]:

$$\alpha \simeq \frac{1}{m}YZ, \tag{a.6}$$

$$\beta \simeq \frac{1}{n}Y'XC'. \tag{a.7}$$

Thus, from (a.4) we get:

$$\begin{aligned}
\hat{C} &\simeq (\alpha'\alpha)^{-1}\alpha'Y\beta(C')^{-1}[C^{-1}\beta'\beta(C')^{-1}]^{-1} \\
&= (\alpha'\alpha)^{-1}\alpha'Y\beta(\beta'\beta)^{-1}C \\
&= (\alpha'\alpha)^{-1}\alpha'(\alpha\beta'+\varepsilon)\beta(\beta'\beta)^{-1}C \\
&= C + (\alpha'\alpha)^{-1}\alpha'\varepsilon\beta(\beta'\beta)^{-1}C \\
&= C + (\frac{1}{n}\alpha'\alpha)^{-1}\frac{1}{mn}\alpha'\varepsilon\beta(\frac{1}{m}\beta'\beta)^{-1}C \\
&\simeq C + [E(\alpha_i\alpha_i')]^{-1}E(\alpha_i\varepsilon_{ij}\beta_j')[E(\beta_j\beta_j')]^{-1}C = C,
\end{aligned}$$

since $E(\alpha_i\varepsilon_{ij}\beta_j') = E(\alpha_i)E(\varepsilon_{ij})E(\beta_j') = 0$, under Assumptions A.1-A.2.

A.1.4 First-order expansion of $\hat{C}$ [Proof of Proposition 3 i)]

The asymptotic equivalences obtained in Appendix 1.3 suggest how to derive the first-order expansion of $\hat{C}$. Let us denote $\beta^* = \beta(C')^{-1}$. We have :

$$Y = \alpha C\beta^{*'} + \varepsilon. \tag{a.8}$$

From (2.6) we also have estimators of α and β^* such that :

$$\sqrt{m}(\hat{\alpha}_i - \alpha_i) \text{ and } \sqrt{n}(\hat{\beta}_j^* - \beta_j^*),$$

are of order $O_p(1)$ (see Lemma A.1 below).

The double IV estimator of C is:

$$\begin{aligned}\hat{C} &= (\hat{\alpha}'\hat{\alpha})^{-1}\hat{\alpha}'Y\hat{\beta}^*(\hat{\beta}^{*'}\hat{\beta}^*)^{-1} \\ &= (\hat{\alpha}'\hat{\alpha})^{-1}\hat{\alpha}'(\alpha C\beta^{*'}+\varepsilon)\hat{\beta}^*(\hat{\beta}^{*'}\hat{\beta}^*)^{-1},\end{aligned}$$

where $\hat{\alpha}=YZ/m$ and $\hat{\beta}^*=Y'X/n$, or equivalently:

$$\begin{aligned}nm(\hat{C}-C) &= \left(\frac{1}{n}\hat{\alpha}'\hat{\alpha}\right)^{-1}\Big\{\hat{\alpha}'[(\alpha-\hat{\alpha})C\hat{\beta}^{*'}+(\alpha-\hat{\alpha})C(\beta^*-\hat{\beta}^*)' \\ &\quad +\hat{\alpha}C(\beta^*-\hat{\beta}^*)'+\varepsilon]\hat{\beta}^*\Big\}(\frac{1}{m}\hat{\beta}^{*'}\hat{\beta}^*)^{-1}.\end{aligned}$$

The terms within the curly brackets have different orders, which are respectively:

$$\begin{aligned}&\frac{nm}{\sqrt{m}}, \text{ for } \hat{\alpha}'(\alpha-\hat{\alpha})C\hat{\beta}^{*'}\hat{\beta}^*, \\ &\frac{nm}{\sqrt{n}\sqrt{m}}, \text{ for } \hat{\alpha}'(\alpha-\hat{\alpha})C(\beta^*-\hat{\beta}^*)'\hat{\beta}^*, \\ &\frac{nm}{\sqrt{n}}, \text{ for } \hat{\alpha}'\hat{\alpha}C(\beta^*-\hat{\beta}^*)'\hat{\beta}^*, \\ &\frac{nm}{\sqrt{n}\sqrt{m}}, \text{ for } \hat{\alpha}'\varepsilon\hat{\beta}^*.\end{aligned}$$

We see that the error-in-variables on α and β^* create the dominant terms in the expansion. We deduce that:

$$\begin{aligned}\sqrt{\min(n,m)}(\hat{C}-C) &= [E(\alpha_i\alpha_i')]^{-1}\{\frac{1}{n}\alpha'[\sqrt{\min(n,m)}(\alpha-\hat{\alpha})]C \\ &\quad +C\frac{1}{m}[\sqrt{\min(n,m)}(\beta^*-\hat{\beta}^*)']\beta^*\}[E(\beta_j^*\beta_j^{*'})]^{-1}+o_P(1). \quad \text{(a.9)}\end{aligned}$$

When n and m tend to infinity at equivalent rates:

$$m=\mu n+o(n), \text{ with } \mu\geq 1, \quad \text{(a.10)}$$

say, we get:

$$\begin{aligned}\sqrt{n}(\hat{C}-C) &= -[E(\alpha_i\alpha_i')]^{-1}\frac{1}{\sqrt{\mu}}[\frac{1}{n}\alpha'\sqrt{m}(\hat{\alpha}-\alpha)]C \\ &\quad -C[\frac{1}{m}\sqrt{n}(\hat{\beta}^*-\beta^*)'\beta^*][E(\beta_j^*\beta_j^{*'})]^{-1}+o_P(1). \quad \text{(a.11)}\end{aligned}$$

To continue this expansion, we need now to derive the expansions of $\sqrt{m}(\hat{\alpha}-\alpha)$ and $\sqrt{n}(\hat{\beta}^*-\beta^*)$.

Lemma A.1: *We have:*

$$\begin{aligned}\sqrt{m}(\hat{\alpha}-\alpha) &= \alpha\{\frac{1}{\sqrt{m}}\sum_{j=1}^{m}[\beta_j z_j' - E(\beta_j z_j')]\} + \frac{1}{\sqrt{m}}\sum_{j=1}^{m}\varepsilon^j z_j',\\ \sqrt{n}(\hat{\beta}^*-\beta^*) &= \beta^* C'\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\alpha_i x_i' - E(\alpha_i x_i')]\} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i x_i'.\end{aligned}$$

Proof: We have:

$$\hat{\alpha}-\alpha = \frac{1}{m}YZ - \alpha = \alpha(\frac{1}{m}\beta' Z - Id) + \frac{1}{m}\varepsilon Z.$$

Therefore:

$$\begin{aligned}\sqrt{m}(\hat{\alpha}-\alpha) &= \alpha\{\sqrt{m}(\frac{1}{m}\beta' Z - Id)\} + \frac{1}{\sqrt{m}}\varepsilon Z\\ &= \alpha\{\frac{1}{\sqrt{m}}\sum_{j=1}^{m}[\beta_j z_j' - E(\beta_j z_j')]\} + \frac{1}{\sqrt{m}}\sum_{j=1}^{m}\varepsilon^j z_j'.\end{aligned}$$

The expansion of $\sqrt{n}(\hat{\beta}^*-\beta^*)$ is obtained with similar arguments.

QED

By applying Lemma A.1, the first term in the RHS of asymptotic expansion (a.11) becomes:

$$\begin{aligned}&-[E(\alpha_i\alpha_i')]^{-1}\frac{1}{\sqrt{\mu}}\{(\frac{1}{n}\alpha'\alpha)\frac{1}{\sqrt{m}}\sum_{j=1}^{m}[\beta_j z_j' - E(\beta_j z_j')] + \frac{1}{n\sqrt{m}}\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_i\varepsilon_{ij}z_j'\}C\\ = \;& -\frac{1}{\sqrt{\mu}}\{\frac{1}{\sqrt{m}}\sum_{j=1}^{m}[\beta_j z_j' - E(\beta_j z_j')]\}C + o_P(1).\end{aligned}$$

By similar arguments, the second term in the RHS of (a.11) is equal to:

$$-C\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[x_i\alpha_i' - E(x_i\alpha_i')]\right\}C + o_P(1),$$

which yields the expansion given in Proposition 3 i).

A.1.5 Asymptotic normality of $\hat{C}$ [Proof of Proposition 3 ii)]

The asymptotic normality is a consequence of the Central Limit Theorem (CLT), which can be applied under Assumptions A.1-A.3. We have just to vectorize the expansion of $\hat{C}$ and then to compute the asymptotic variance-covariance matrix. We have by applying formula (2.8):

$$\begin{aligned}vec[\sqrt{n}(\hat{C}-C)] &= -\frac{1}{\sqrt{\mu}}\left\{(C'\otimes Id)\;vec\left(\frac{1}{\sqrt{m}}\sum_{j=1}^{m}[\beta_j z_j' - E(\beta_j z_j')]\right)\right\}\\ &\quad -\left\{(C'\otimes C])vec\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[x_i\alpha_i' - E(x_i\alpha_i')]\right)\right\} + o_P(1),\end{aligned}$$

and:

$$vec[\sqrt{n}(\hat{C}-C)] = -\frac{1}{\sqrt{\mu}}\left\{(C'\otimes Id)\left(\frac{1}{\sqrt{m}}\sum_{j=1}^{m}[z_j\otimes\beta_j - E(z_j\otimes\beta_j)]\right)\right\} - \left\{(C'\otimes C])vec\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\alpha_i\otimes x_i - E(\alpha_i\otimes x_i)]\right)\right\} + o_P(1),$$

by using that $vec(ab') = b\otimes a$ for a pair of vectors a and b. Proposition 3 ii) follows, since under Assumption A.3, the two components in the first-order expansion for $\hat{C}$ are independent.

A.1.6 First-order expansion of the matrix of fitted values [Proof of Proposition 4 i)]

Let us now consider the expansion of the matrix of fitted values:

$$\hat{Y} = \hat{\alpha}\hat{\beta}' = (\frac{1}{m}YZ)\hat{C}(\frac{1}{n}X'Y).$$

We have seen in Section 1.5 of the Appendix that the first-order expansion of $\hat{C}$ does not involve the effect of error terms ε. Therefore, in deriving the expansion of $\hat{Y}$, we can replace Y by $\alpha\beta'$ in equation (2.10). We get:

$$\begin{aligned}\hat{C} &\simeq (\frac{1}{m^2}Z'\beta\alpha'\alpha\beta'Z)^{-1}(\frac{1}{m}Z'\beta\alpha')\alpha\beta'(\frac{1}{n}\beta\alpha'X)(\frac{1}{n^2}X'\alpha\beta'\beta\alpha'X)^{-1}\\ &= (\frac{1}{m}\beta'Z)^{-1}(\alpha'\alpha)^{-1}(\frac{1}{m}Z'\beta)^{-1}(\frac{1}{m}Z'\beta)\alpha'\alpha\beta'\beta(\frac{1}{n}\alpha'X)(\frac{1}{n}\alpha'X)^{-1}(\beta'\beta)^{-1}(\frac{1}{n}X'\alpha)^{-1}\\ &= (\frac{1}{m}\beta'Z)^{-1}(\frac{1}{n}X'\alpha)^{-1}.\end{aligned}$$

We deduce:

$$\begin{aligned}\hat{Y} &\simeq (\frac{1}{m}YZ)(\frac{1}{m}\beta'Z)^{-1}(\frac{1}{n}X'\alpha)^{-1}(\frac{1}{n}X'Y)\\ &= (\frac{1}{m}\alpha\beta'Z+\frac{1}{m}\varepsilon Z)(\frac{1}{m}\beta'Z)^{-1}(\frac{1}{n}X'\alpha)^{-1}(\frac{1}{n}X'\alpha\beta'+\frac{1}{n}X'\varepsilon)\\ &\simeq \alpha\beta'+\frac{1}{m}\varepsilon Z(\frac{1}{m}\beta'Z)^{-1}\beta'+\alpha(\frac{1}{n}X'\alpha)^{-1}\frac{1}{n}X'\varepsilon\\ &\simeq \alpha\beta'+\frac{1}{m}\varepsilon Z[E(\beta_j z_j')]^{-1}\beta'+\alpha[E(x_i\alpha_i')]^{-1}\frac{1}{n}X'\varepsilon.\end{aligned}$$

We deduce that:

$$\sqrt{n}(\hat{Y}-\alpha\beta') = \frac{1}{\sqrt{\mu}}(\frac{1}{\sqrt{m}}\varepsilon Z)[E(\beta_j z_j')]^{-1}\beta'+\alpha[E(x_i\alpha_i')]^{-1}(\frac{1}{\sqrt{n}}X'\varepsilon)+o_P(1).$$

By applying the vec operator and formula (2.8), we get :

$$\begin{aligned} vec[\sqrt{n}(\hat{Y}-\alpha\beta')] &= \frac{1}{\sqrt{\mu}}\{(\beta[E(z_j\beta_j')]^{-1})\otimes Id_n\}vec(\frac{1}{\sqrt{m}}\varepsilon Z) \\ &\quad +\{Id_m\otimes(\alpha[E(x_i\alpha_i')]^{-1})\}vec(\frac{1}{\sqrt{n}}X'\varepsilon)+o_P(1). \end{aligned}$$

A.1.7 Asymptotic normality of the matrix of fitted values [Proof of Proposition 4 ii)]

The asymptotic normality is a consequence of the CLT. We have just to derive the asymptotic variance. By Assumption A.3, $vec(\frac{1}{\sqrt{m}}\varepsilon Z)$ and $vec(\frac{1}{\sqrt{n}}X'\varepsilon)$ are asymptotically non correlated. Moreover,

$$\begin{aligned} V_{as}[vec(\frac{1}{\sqrt{n}}X'\varepsilon)] &= V[vec(x_i\varepsilon_i')] = V(\varepsilon_i\otimes x_i) \\ &= V\begin{pmatrix} x_i\varepsilon_{i1} \\ \vdots \\ x_i\varepsilon_{im}\end{pmatrix} = E\left\{V\left[\begin{pmatrix} x_i\varepsilon_{i1} \\ \vdots \\ x_i\varepsilon_{im}\end{pmatrix}|x_i\right]\right\} \\ &= Id_m\otimes[\sigma^2E(x_ix_i')]. \end{aligned}$$

Similarly:

$$V_{as}[vec(\frac{1}{\sqrt{m}}\varepsilon Z)] = [\sigma^2E(z_jz_j')]\otimes Id_n.$$

The result in Proposition 4 ii) follows.

Appendix 2: Analysis of social distances

Model (2.1) is a pure descriptive model. But the introduction of appropriate restrictions can make this model more structural. As an illustration, let us consider the determination of social distances between individuals from observations $y_{i,j}$ of their joint decisions. In this framework the matrix of observations is symmetric, and the model can be written as :

$$y_{i,j} = c + (\beta_i-\beta_j)'\Omega(\beta_i-\beta_j)+\varepsilon_{i,j}, \qquad \text{(a.12)}$$

where β_i is an unobserved vector of attributes, which are i.i.d zero-mean, [12] Ω an unknown symmetric matrix, c an unknown scalar, and the errors are such that $\varepsilon_{i,j} = \varepsilon_{j,i}$ and the components of $vech(\varepsilon)$ are i.i.d. This specification assumes exogenous attributes and does not apply to the case of endogenous attributes such as endogenous choice of the individual position in the network. These attributes characterize the positions of the individuals in the social space. The symmetric matrix Ω can have negative as well as positive eigenvalues. The eigenvectors associated with negative eigenvalues corresponds to attributes satisfying the condition of homophily of attributes; the eigenvectors associated with positive eigenvalues adapt the model to situations where opposites attract. This model is a constrained version of the model discussed in Section 2.3 i). Indeed, we have :

$$y_{i,j} = c + \beta_i'\Omega\beta_i + \beta_j'\Omega\beta_j - 2\beta_i'\Omega\beta_j + \varepsilon_{i,j}. \tag{a.13}$$

We get an ANOVA model the type :

$$y_{i,j} = c + a_i + b_j + d_{i,j} + \varepsilon_{i,j}, \tag{a.14}$$

with both marginal and cross-effects. These effects are constrained, since they depend on a rather small number of latent parameters, that are, the $\beta_i's$, Ω, and c. We can transform the observations as in Section 2.3 i) to get

$$\tilde{y}_{i,j} = y_{i,j} - y_{i,.} - y_{.,j} + y_{.,.} \simeq -2\beta_i'\Omega\beta_j + \varepsilon_{i,j}. \tag{a.15}$$

We get a specification compatible with model (2.1), where:

$$\alpha_i = -2\Omega\beta_i, \tag{a.16}$$

depends on β_i. Hence, factors α_i and β_i are dependent.

This constrained model can be analyzed in two different ways.

i) We can first estimate the unconstrained model to derive $\hat{\alpha}_i, \hat{\beta}_i$ along the lines of Section 2.2 and Proposition 2. Then, in a second step, the constraints can be taken into account by considering the asymptotic least squares estimation [Gourieroux, Monfort, Trognon (1982), Kodde, Palm, Pfann (1990)], that is by minimizing :

[12] The zero-mean assumption can always be made, since the equation depends on the $\beta_i's$ by means of differences only.

$$\min \sum_{i=1}^{n} \{\|\hat{\alpha}_i + 2\Omega\beta_i\|^2 + \|\hat{\beta}_i - \beta_i\|^2\}, \tag{a.17}$$

with respect to $\beta_i, i = 1, \ldots, n$, and Ω under the symmetry restriction. The second-step estimators of α_i, β_i are $\hat{\hat{\alpha}}_i = 2\hat{\Omega}\hat{\hat{\beta}}_i, \hat{\hat{\beta}}_i$, where $\hat{\Omega}, \hat{\hat{\beta}}_i, i = 1, \ldots, n$ are the solutions of minimization problem (a.17).

ii) Alternatively, we can apply an (asymptotic) instrumental variable $\hat{x}_i$, say, and deduce an estimate of $\beta_j, \hat{\beta}_j$, say, up to an invertible (K, K) matrix. Then in a second step, we can regress by OLS $y_{i,j}$ on $\hat{\beta}_i'\Omega\hat{\beta}_j$ to deduce a consistent estimator $\hat{\Omega}$ of Ω.

Anyway the method provides both consistent approximations of the unobserved attributes and select the appropriate dissimilarity measure Ω. Thus, no arbitrary choice of dissimilarity is required.

Appendix 3: Principal Component Analysis (PCA)

In this Appendix, we first briefly review Principal Component Analysis (PCA) and its interpretations [see e.g. Jolliffe (2002) for a textbook presentation]. Then, we derive the large sample properties of the matrix of fitted values obtained from PCA.

A.3.1 PCA and its interpretations

In this subsection we consider PCA in terms of either the joint spectral analysis of matrices YY' and $Y'Y$, or as an estimator derived by the Least Squares (LS) principle, or as an estimator obtained from the Expectation Maximization (EM) algorithm in the limiting case of vanishing noise.

i) Spectral decomposition of matrices $Y'Y$ and YY'

Let $Y = (y_{i,j})$ be a (n, m) stochastic matrix. Let us consider the transposed rows y_i, $i =$

$1, ..., n$, of matrix Y as a sample of n observations of a m-dimensional random vector. For expository purpose, let us assume that the data have been centered such that vectors y_i have zero sample mean. In PCA we look for $K \leq \min\{n, m\}$ orthonormal vectors b_1, ..., b_K in $\mathbf{R}^m$, such that $b_1' y_i$ has the largest sample variance, $b_2' y_i$ has the largest sample variance under the constraint that b_2 is orthogonal to b_1, and so on. The solution to this problem consists of the normalized eigenvectors of the (m, m) sample variance-covariance matrix $\frac{1}{n} Y'Y = \frac{1}{n} \sum_{i=1}^{n} y_i y_i'$ associated with the K largest eigenvalues. Then, the (m, K) matrix $\hat{b} = [\hat{b}_1, ..., \hat{b}_K]$ is such that:

$$\frac{1}{n} Y'Y \hat{b} = \hat{b} \hat{\Lambda}, \qquad \hat{b}' \hat{b} = Id_K, \tag{a.18}$$

where $\hat{\Lambda}$ is the diagonal matrix of the K largest eigenvalues of $Y'Y/n$. The (sample) Principal Components are the columns of the (n, K) matrix $\hat{a}$ defined by

$$\hat{a} = Y \hat{b}. \tag{a.19}$$

They correspond to K linear aggregates of the data that retain the maximal variability.

The (n, n) matrix $\frac{1}{m} YY' = \frac{1}{m} \sum_{j=1}^{m} y^j (y^j)'$ is the sample variance-covariance matrix of the n-dimensional columns of Y. Matrices YY' and $Y'Y$ have the same non zero eigenvalues. In particular, we have:

$$\frac{1}{n} YY' \hat{a} = Y \frac{1}{n} Y'Y \hat{b} = Y \hat{b} \hat{\Lambda} = \hat{a} \hat{\Lambda},$$

and the columns of matrix $\hat{a}$ are eigenvectors of YY' associated with the K largest eigenvalues.

ii) Least Squares principle

Let us consider the factor model $y_{i,j} = \alpha_i' \beta_j + \varepsilon_{i,j}$, or in matrix format $Y = \alpha \beta' + \varepsilon$, where the factors α and β satisfy the assumptions introduced in Section 2. Let us further adopt the standard identification restriction $E(\beta_j \beta_j') = Id_K$. This identification restriction leaves free a rotation and sign changes of the factors. The rotation can be fixed by the restriction that matrix $E(\alpha_i \alpha_i')$ is diagonal.

If we treat the factor values α and β as "parameters" (fixed row and column effects), we can define estimators of α and β by minimizing the Least Squares (LS) criterion:

$$\min_{\{\alpha_i\}, \{\beta_j\}} \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{i,j} - \alpha_i' \beta_j)^2,$$

subject to the sample identification restriction:

$$\frac{1}{m}\sum_{j=1}^{m}\beta_j\beta_j' = Id_K.$$

Equivalently in matrix notation, we get the constrained minimization problem:

$$\min_{\alpha,\beta} Tr[(Y-\alpha\beta')'(Y-\alpha\beta')] \qquad \text{(a.20)}$$
$$s.t.\ \beta'\beta/m = Id_K.$$

The criterion and the constraint are invariant to factor rotations, i.e. to mappings $\beta \to \beta C$, $\alpha \to \alpha C$ where C is an orthogonal (K,K) matrix. We can fix this rotational invariance by the identification restriction that matrix $\alpha'\alpha$ is diagonal.

The F.O.C. of problem (a.20) yield:

$$\alpha = Y\beta(\beta'\beta)^{-1}, \qquad \text{(a.21)}$$
$$\beta = Y'\alpha(\alpha'\alpha)^{-1}, \qquad \text{(a.22)}$$

and the Lagrange multipliers for the matrix restriction $\beta'\beta/m = Id_K$ (and $\alpha'\alpha$ diagonal) vanish. The set of solutions of the nonlinear system of equations (a.21)-(a.22) is stable under mapping $\beta \to \beta C$, $\alpha \to \alpha(C')^{-1}$, where C is a nonsingular (K,K) matrix. The constraint $\beta'\beta/m = Id_K$ restricts the set of invariant transformations to orthogonal matrices. The estimator is a solution of the nonlinear system of equations (a.21)-(a.22) satisfying the identification restrictions $\beta'\beta/m = Id_K$ and $\alpha'\alpha$ diagonal.

By plugging (a.21) into (a.20), the criterion concentrated w.r.t. α becomes:

$$\begin{aligned} Tr[(Y-\alpha\beta')'(Y-\alpha\beta)] &= Tr(M_\beta Y'Y M_\beta) = Tr(Y'Y M_\beta) \\ &= Tr(Y'Y) - Tr[\beta'(\frac{1}{m}Y'Y)\beta], \end{aligned}$$

where $M_\beta = Id_m - \beta(\beta'\beta)^{-1}\beta' = Id_m - \beta\beta'/m$ from the identification restriction $\beta'\beta/m = Id_K$, and we use the commutative property of the trace operator. Thus, after neglecting irrelevant additive terms and rescaling the criterion, the optimization problem defining the estimator becomes:

$$\max_{\beta} Tr[\beta'(\frac{1}{mn}Y'Y)\beta]$$
$$s.t.\ \beta'\beta/m = Id_K.$$

A solution $\hat{\beta}$ of this optimization problem is the matrix of the normalized eigenvectors associated with the K largest eigenvalues of the symmetric (m,m) matrix $Y'Y/(mn)$:

$$(\frac{1}{mn}Y'Y)\hat{\beta} = \hat{\beta}\hat{D}, \quad \hat{\beta}'\hat{\beta}/m = Id_K, \tag{a.23}$$

where $\hat{D}$ is the (K,K) diagonal matrix of the eigenvalues. From (a.21) and the identification restriction, the estimator of α is:

$$\hat{\alpha} = Y\hat{\beta}/m. \tag{a.24}$$

The rotation fix that is selected is such that the matrix $\hat{\alpha}'\hat{\alpha}/n$ is diagonal, equal to $\hat{D}$.

By comparing equations (a.18)-(a.19) and (a.23)-(a.24), estimators $\hat{\alpha}$ and $\hat{\beta}$ are rescaled versions of matrices $\hat{a}$ and $\hat{b}$ obtained from PCA in paragraph i):

$$\hat{\alpha} = \frac{1}{\sqrt{m}}\hat{a}, \quad \hat{\beta} = \sqrt{m}\hat{b},$$

and $\hat{D} = \hat{\Lambda}/m$.

iii) Maximum Likelihood (ML) and Expectation Maximization (EM) algorithm

The PCA estimator can also be interpreted from the view point of the EM algorithm [see e.g. Chen (2002), Section 3]. Let us write the factor model for the transposed rows of matrix Y as $y_i = \beta\alpha_i + \varepsilon_i$, with $i = 1, ..., n$, and assume independent Gaussian distributions for the latent factor $\alpha_i \sim IIN(0, Id_K)$ and the errors $\varepsilon_i \sim IIN(0, \sigma^2 Id_m)$. The scalar $\sigma^2 > 0$ and the (m,K) matrix β are unknown parameters. We adopt the identification restriction that matrix $\beta'\beta$ is diagonal.

The distribution of vector y_i is Gaussian $y_i \sim N(0, \beta\beta' + \sigma^2 Id_m)$. We get the log-likelihood function:

$$\begin{aligned}
\mathcal{L}(\beta, \sigma^2) &= \frac{1}{n}\sum_{i=1}^{n} \log f(y_i|\beta, \sigma^2) \\
&= -\frac{1}{2}\log\det\Omega(\beta,\sigma^2) - \frac{1}{2n}\sum_{i=1}^{n} y_i'\Omega(\beta,\sigma^2)^{-1}y_i,
\end{aligned}$$

where $\Omega(\beta, \sigma^2) = \beta\beta' + \sigma^2 Id_m$. By maximizing this function w.r.t. parameters β and σ^2 such that $\beta'\beta$ is diagonal, we get the estimators:

$$\begin{aligned}
\hat{\beta}_k &= (\hat{\lambda}_k - \hat{\sigma}^2)^{1/2}\hat{b}_k, \quad k = 1, ..., K, \\
\hat{\sigma}^2 &= [Tr(Y'Y/n) - \sum_{k=1}^{K}\hat{\lambda}_k]/(m-K),
\end{aligned}$$

where the $\hat{\beta}_k$ denote the columns of the estimated matrix $\hat{\beta}$, and the $\hat{b}_k$ are the orthonormal eigenvectors of matrix $Y'Y/n$ associated to the K largest eigenvalues $\hat{\lambda}_k$.

The ML estimate can also be computed by the EM algorithm. The full-observation log-density of Y and α is given by:

$$\log f(Y, \alpha|\beta, \sigma^2) = -\frac{m}{2}\log\sigma^2 - \frac{1}{2n\sigma^2}\sum_{i=1}^{n}(y_i - \beta\alpha_i)'(y_i - \beta\alpha_i),$$

up to terms that do not depend on parameters β and σ^2. This log-density has to be integrated w.r.t. the conditional distribution of the latent factor α conditional on observations Y. By the joint normality, we have:

$$\alpha_i|y_i \sim N\left(\beta'(\beta\beta' + \sigma^2 Id_m)^{-1}y_i, Id_K - \beta'(\beta\beta' + \sigma^2 Id_m)^{-1}\beta\right).$$

For β such that $\beta'\beta$ is diagonal, we have $\beta'(\beta\beta' + \sigma^2 Id_m)^{-1} = (\beta'\beta + \sigma^2 Id_K)^{-1}\beta'$. Then, we get:

$$\alpha_i|y_i \sim N\left((\beta'\beta + \sigma^2 Id_K)^{-1}\beta'y_i, \sigma^2(\beta'\beta + \sigma^2 Id_K)^{-1}\right).$$

The mean of this Gaussian distribution is given by the Ridge regression of the data vector y_i on the "regressor matrix" β. When either m gets large such that the diagonal elements of $\beta'\beta/m$ are strictly positive, or σ^2 gets small, the variance of the Gaussian distribution shrinks to zero, and the conditional distribution peaks at $(\beta'\beta)^{-1}\beta'y_i$.

Let us now consider the Expectation (E) and Maximization (M) steps of the iterative algorithm. Let $\tilde{\beta}$ and $\tilde{\sigma}^2$ be estimates from the previous iteration. In the E-step, we compute the expectation of $\log f(Y, \alpha|\beta, \sigma^2)$ w.r.t. the distribution of α given Y for parameters $\tilde{\beta}$ and $\tilde{\sigma}^2$, to get the function:

$$\mathcal{Q}(\beta, \sigma^2|\tilde{\beta}, \tilde{\sigma}^2) = \underset{\tilde{\beta},\tilde{\sigma}^2}{E}[\log f(Y, \alpha|\beta, \sigma^2)|Y].$$

Define $\hat{\alpha}_i = (\tilde{\beta}'\tilde{\beta} + \tilde{\sigma}^2 Id_K)^{-1}\tilde{\beta}'y_i$ and:

$$\hat{\alpha} = Y\tilde{\beta}(\tilde{\beta}'\tilde{\beta} + \tilde{\sigma}^2 Id_K)^{-1}, \qquad \hat{\Sigma} = \tilde{\sigma}^2(\tilde{\beta}'\tilde{\beta} + \tilde{\sigma}^2 Id_K)^{-1}. \tag{a.25}$$

We have:

$$\begin{aligned}
\mathcal{Q}(\beta, \sigma^2|\tilde{\beta}, \tilde{\sigma}^2) &= -\frac{m}{2}\log\sigma^2 - \frac{1}{2n\sigma^2}\sum_{i=1}^{n}(y_i - \beta\hat{\alpha}_i)'(y_i - \beta\hat{\alpha}_i) - \frac{1}{2\sigma^2}Tr(\beta\hat{\Sigma}\beta') \\
&= -\frac{m}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(\frac{1}{n}Tr[(Y - \hat{\alpha}\beta')'(Y - \hat{\alpha}\beta')] + Tr(\beta\hat{\Sigma}\beta')\right).
\end{aligned}$$

In the M-step, function $\mathcal{Q}(\beta, \sigma^2|\tilde{\beta}, \tilde{\sigma}^2)$ is maximized w.r.t. the parameters β and σ^2. We have:

$$\begin{aligned} & \frac{1}{n}Tr[(Y-\hat{\alpha}\beta')'(Y-\hat{\alpha}\beta')] + Tr(\beta\hat{\Sigma}\beta') \\ = \; & Tr\left[\left(\beta-(Y'\hat{\alpha}/n)(\hat{\alpha}'\hat{\alpha}/n+\hat{\Sigma})^{-1}\right)(\hat{\alpha}'\hat{\alpha}/n+\hat{\Sigma})\left(\beta-(Y'\hat{\alpha}/n)(\alpha'\hat{\alpha}/n+\hat{\Sigma})^{-1}\right)'\right] \\ & +\frac{1}{n}Tr\left[Y'\left(Id_n-\frac{1}{n}\hat{\alpha}(\hat{\alpha}'\hat{\alpha}/n+\hat{\Sigma})^{-1}\hat{\alpha}'\right)Y\right]. \end{aligned}$$

Thus, the minimizer of function $(\beta, \sigma^2) \to \mathcal{Q}(\beta, \sigma^2|\tilde{\beta}, \tilde{\sigma}^2)$ is:

$$\begin{aligned} \hat{\beta} &= (Y'\hat{\alpha}/n)(\hat{\alpha}'\hat{\alpha}/n+\hat{\Sigma})^{-1}, \\ \hat{\sigma}^2 &= \frac{1}{nm}Tr\left[Y'\left(Id_n-\frac{1}{n}\hat{\alpha}(\hat{\alpha}'\hat{\alpha}/n+\hat{\Sigma})^{-1}\hat{\alpha}'\right)Y\right]. \end{aligned} \tag{a.26}$$

Equations (a.25) and (a.26) define the updating rule for the parameters.

When m is large, and the iterations remain in the region of the parameter space such that $\beta'\beta/m$ and $\alpha'\alpha/n$ are positive definite matrices, matrix $\hat{\Sigma}$ is close to zero and the updating rule becomes:

$$\hat{\alpha} = Y\tilde{\beta}(\tilde{\beta}'\tilde{\beta})^{-1}, \tag{a.27}$$

$$\hat{\beta} = Y'\hat{\alpha}(\hat{\alpha}'\hat{\alpha})^{-1}. \tag{a.28}$$

These equations correspond to an iterative algorithm for solving the FOC in (a.21)-(a.22).

A.3.2 Asymptotic distribution of the matrix of fitted values from PCA

Let us consider the factor model $y_{i,j} = \alpha_i'\beta_j + \varepsilon_{i,j}$, where the K-dimensional latent factors α_i and β_j and the errors $\varepsilon_{i,j}$ satisfy the assumptions in Section 2. The matrix of fitted values $\alpha\beta'$ is invariant under one-to-one mappings of the factor space such that $\alpha \to \alpha C$ and $\beta \to \beta(C^{-1})'$, where C is a non-singular (K, K) matrix. Thus, parameter $\alpha\beta'$ does not depend on the selected identification restrictions. It is convenient to adopt the restrictions $E(\beta_j\beta_j') = Id_K$ and $E(\alpha_i\alpha_i') = D$ diagonal. We assume that the (unknown) diagonal elements of matrix $D = diag(d_1, ..., d_K)$ are strictly positive, distinct and ranked in descending order: $d_1 > d_2 > ... > d_K > 0$.

i) Asymptotic expansions of estimators $\hat{\alpha}$ and $\hat{\beta}$

The PCA estimators $\hat{\alpha}$ and $\hat{\beta}$ are defined by equations (a.23) and (a.24). We derive asymptotic expansions of these PCA estimators. We build on the analysis provided in e.g. Bai, Ng (2002),

Stock, Watson (2002), and modify some steps of the derivation in order to obtain a novel result on the asymptotic distribution of the fitted values (Proposition 7).

The next lemma shows that the diagonal matrix $\hat{D}$ providing the K largest eigenvalues of $Y'Y/(nm)$ ranked in descending order converges in probability to D [see Section iii) below for the proof]. The eigenvalues of matrix $Y'Y/(mn)$ can be ranked, because the continuous distributions of α, β, ε ensure that the probability of equality of two eigenvalues is zero.

Lemma A.2: *As $n, m \to \infty$ such that $m = \mu n + o(n)$, with $\mu \geq 1$, we have: $\hat{D} \xrightarrow{p} D$.*

From Lemma A.2, matrix $\hat{D}$ is invertible with probability approaching (w.p.a.) 1. Then, by using equation (a.23) and replacing $Y = \alpha\beta' + \varepsilon$, we get:

$$\begin{aligned}\hat{\beta} &= (\frac{1}{nm}Y'Y)\hat{\beta}\hat{D}^{-1} \\ &= \beta(\frac{\alpha'\alpha}{n})(\frac{\beta'\hat{\beta}}{m})\hat{D}^{-1} + \frac{1}{n}\varepsilon'\alpha(\frac{\beta'\hat{\beta}}{m})\hat{D}^{-1} + \beta\frac{1}{mn}\alpha'\varepsilon\hat{\beta}\hat{D}^{-1} + \frac{1}{mn}\varepsilon'\varepsilon\hat{\beta}\hat{D}^{-1}. \qquad \text{(a.29)}\end{aligned}$$

From equation (a.24) using $Y = \alpha\beta' + \varepsilon$, we get:

$$\hat{\alpha} = \alpha(\frac{\beta'\hat{\beta}}{m}) + \frac{1}{m}\varepsilon\hat{\beta}.$$

By replacing $\hat{\beta}$ in the second term of the r.h.s. using equation (a.29), we get:

$$\begin{aligned}\hat{\alpha} &= \alpha(\frac{\beta'\hat{\beta}}{m}) + \frac{1}{m}\varepsilon\beta(\frac{\alpha'\alpha}{n})(\frac{\beta'\hat{\beta}}{m})\hat{D}^{-1} + \frac{1}{mn}\varepsilon\varepsilon'\alpha(\frac{\beta'\hat{\beta}}{m})\hat{D}^{-1} \\ &\quad + \frac{1}{m^2n}\varepsilon\beta\alpha'\varepsilon\hat{\beta}\hat{D}^{-1} + \frac{1}{m^2n}\varepsilon\varepsilon'\varepsilon\hat{\beta}\hat{D}^{-1} \qquad \text{(a.30)}\end{aligned}$$

Let us now show that matrix $\beta'\hat{\beta}/m$ converges in probability to the identity matrix Id_K, up to the sign choice for the estimated factors. All terms in the r.h.s. of equation (a.29) vanish asymptotically except the first one. Thus, we have:

$$\begin{aligned}Id_K &= \frac{\hat{\beta}'\hat{\beta}}{m} = \hat{D}^{-1}(\frac{\beta'\hat{\beta}}{m})'(\frac{\alpha'\alpha}{n})(\frac{\beta'\beta}{m})(\frac{\alpha'\alpha}{n})(\frac{\beta'\hat{\beta}}{m})\hat{D}^{-1} + o_p(1) \\ &= D^{-1}(\frac{\beta'\hat{\beta}}{m})'D^2(\frac{\beta'\hat{\beta}}{m})D^{-1} + o_p(1), \qquad \text{(a.31)}\end{aligned}$$

where we use Lemma A.2. Similarly, from equation (a.30) we have:

$$\hat{D} = \frac{\hat{\alpha}'\hat{\alpha}}{n} = (\frac{\beta'\hat{\beta}}{m})'(\frac{\alpha'\alpha}{n})(\frac{\beta'\hat{\beta}}{m}) + o_p(1) = (\frac{\beta'\hat{\beta}}{m})'D(\frac{\beta'\hat{\beta}}{m}) + o_p(1),$$

which implies from Lemma A.2:

$$D = (\frac{\beta'\hat{\beta}}{m})'D(\frac{\beta'\hat{\beta}}{m}) + o_p(1). \tag{a.32}$$

Let us now define matrix $\hat{Q} = D^{1/2}(\frac{\beta'\hat{\beta}}{m})D^{-1/2}$. Then, equations (a.31) and (a.32) imply that matrix $\hat{Q}$ is such that:

$$\hat{Q}'\hat{Q} = Id_K + o_p(1), \quad \hat{Q}'D\hat{Q} = D + o_p(1).$$

The next lemma is proved in Section iii) below.

Lemma A.3: *Let $\hat{Q}$ be a square (K, K) stochastic matrix such that $\hat{Q}'\hat{Q} = Id_K + o_p(1)$ and $\hat{Q}'D\hat{Q} = D + o_p(1)$ as $n, m \to \infty$, where $D = diag(d_1, ..., d_K)$ is a diagonal matrix with diagonal elements $d_1 > d_2 > ... > d_K > 0$. Then $\hat{Q} = \hat{S} + o_p(1)$ as $n, m \to \infty$, where $\hat{S}$ is a diagonal matrix such that $\hat{S}^2 = Id_K$.*

From Lemma A.3 we have $\hat{Q} = \hat{S} + o_p(1)$, and thus $\beta'\hat{\beta}/m = \hat{S} + o_p(1)$. In particular, matrix $\beta'\hat{\beta}/m$ is invertible w.p.a. 1. The diagonal elements of the stochastic matrix $\hat{S}$ are either 1, or -1, and correspond to the sign indeterminacy of the factor estimates. Without loss of generality we can assume that $\hat{S} = Id_K$.

By the invertibility of matrix $\beta'\hat{\beta}/m$, and equation (a.30), we have $\alpha \simeq \hat{\alpha}(\frac{\beta'\hat{\beta}}{m})^{-1}$, neglecting higher order terms. Thus:

$$\frac{\alpha'\alpha}{n} \simeq [(\frac{\beta'\hat{\beta}}{m})']^{-1}\hat{D}(\frac{\beta'\hat{\beta}}{m})^{-1}.$$

By using this expansion in equations (a.29) and (a.30), as well as $\beta'\hat{\beta}/m = Id_K + o_p(1)$ and $\hat{D}^{-1} = D^{-1} + o_p(1)$, we get the asymptotic expansions:

$$\hat{\beta} \simeq \beta[(\frac{\beta'\hat{\beta}}{m})']^{-1} + \frac{1}{n}\varepsilon'\alpha D^{-1}, \tag{a.33}$$

$$\hat{\alpha} \simeq \alpha(\frac{\beta'\hat{\beta}}{m}) + \frac{1}{m}\varepsilon\beta. \tag{a.34}$$

ii) Asymptotic distribution of the fitted values (proof of Proposition 7)

The fitted values matrix from PCA is $\hat{Y}^{PCA} = \hat{\alpha}\hat{\beta}'$. From the asymptotic expansions (a.33) and (a.34), we get:

$$\hat{Y}^{PCA} \simeq \alpha\beta' + \alpha(\beta'\hat{\beta}/m)D^{-1}(\alpha'\varepsilon/n) + (\varepsilon\beta/m)(\beta'\hat{\beta}/m)^{-1}\beta'.$$

Then, using $\beta'\hat{\beta}/m = Id_K + o_p(1)$ we get:

$$\sqrt{n}(\hat{Y}^{PCA} - \alpha\beta') \simeq \alpha D^{-1}\left(\frac{\alpha'\varepsilon}{\sqrt{n}}\right) + \frac{1}{\sqrt{\mu}}\left(\frac{\varepsilon\beta}{\sqrt{m}}\right)\beta'.$$

By comparing with Proposition 4, and recalling that $E(\alpha_i\alpha_i') = D$ and $E(\beta_j\beta_j') = Id_K$, we deduce that the PCA estimator of the fitted values $\hat{Y}^{PCA}$ is asymptotically equivalent to the (unfeasible) double IV estimator based on instruments $x_i = \alpha_i$ and $z_j = \beta_j$. The asymptotic variance is given by:

$$V_{as}[vec(\sqrt{n}(\hat{Y}^{PCA} - \alpha\beta'))] = \frac{\sigma^2}{\mu}\left\{\beta E[\beta_j\beta_j']^{-1}\beta'\right\} \otimes I_n + \sigma^2 Id_m \otimes \left\{\alpha E[\alpha_i\alpha_i']^{-1}\alpha'\right\}.$$

As in Proposition 4, the dimensions (N, M) of the matrix of fitted values are kept constant in the asymptotics.

iii) Proofs of the lemmas

Proof of Lemma A.2: The proof uses the singular value version of the Weyl's inequalities [Horn and Johnson (1985), Theorem 3.3.16]. Let $\lambda_k(\cdot)$ denote the k-th largest eigenvalue of a symmetric matrix. For (n, m) matrices A and B, we have the following inequalities on the square roots of the ranked eigenvalues of matrices $A'A$, $B'B$ and $(A+B)'(A+B)$:

$$[\lambda_k((A+B)'(A+B))]^{1/2} \leq [\lambda_k(A'A)]^{1/2} + [\lambda_1(B'B)]^{1/2},$$

and:

$$[\lambda_k((A+B)'(A+B))]^{1/2} \geq [\lambda_k(A'A)]^{1/2} - [\lambda_1(B'B)]^{1/2},$$

for any k such that $1 \leq k \leq \min\{n, m\}$. Now, $\hat{D}$ is the diagonal matrix of the K largest eigenvalues of $Y'Y/(nm)$, and $Y = \alpha\beta' + \varepsilon$. From the singular value version of the Weyl's inequalities, it follows:

$$\left|\left[\lambda_k\left(\frac{1}{mn}Y'Y\right)\right]^{1/2} - \left[\lambda_k\left(\frac{1}{m}\beta(\frac{\alpha'\alpha}{n})\beta'\right)\right]^{1/2}\right| \leq \lambda_1\left[\left(\frac{1}{mn}\varepsilon'\varepsilon\right)\right]^{1/2}, \qquad \text{(a.35)}$$

for any $k = 1, ..., K$. We have that $\lambda_1(\varepsilon'\varepsilon/n)$ converges almost surely to $(1+\sqrt{\mu})^2\sigma^2$ as $n, m \to \infty$ such that $m/n \to \mu$, if the errors $\varepsilon_{i,j}$ have finite fourth-order moments [see e.g. Geman (1980) and Yin, Bai, Krishnaiah (1988)]. Thus, we get:

$$\lambda_k\left(\frac{1}{mn}Y'Y\right) = \lambda_k\left(\frac{1}{m}\beta(\frac{\alpha'\alpha}{n})\beta'\right) + o_p(1), \qquad \text{(a.36)}$$

for any $k = 1, ..., K$.

The symmetric, positive semi-definite matrix $A = \frac{1}{m}\beta(\frac{\alpha'\alpha}{n})\beta'$ has rank K, and its non-zero eigenvalues correspond to eigenvectors that are in the column space of matrix β. An orthonormal basis of this column space is provided by the columns of matrix $Z = \frac{1}{\sqrt{m}}\beta(\beta'\beta/m)^{-1/2}$. Thus, the K largest eigenvalues of matrix A are the eigenvalues of the (K, K) matrix $Z'AZ = (\beta'\beta/m)^{1/2}(\alpha'\alpha/n)(\beta'\beta/m)^{1/2}$. This matrix converges to D in probability. By the continuity of the matrix eigenvalue function $\lambda_k(\cdot)$, and the fact that the diagonal matrix D has distinct diagonal elements, we get:

$$\begin{aligned} \lambda_k\left(\frac{1}{m}\beta(\frac{\alpha'\alpha}{n})\beta'\right) &= \lambda_k\left((\beta'\beta/m)^{1/2}(\alpha'\alpha/n)(\beta'\beta/m)^{1/2}\right) \\ &= \lambda_k(D) + o_p(1) = d_k + o_p(1). \end{aligned} \tag{a.37}$$

From (a.36) and (a.37), the conclusion follows. Q.E.D.

Proof of Lemma A.3: The condition $\hat{Q}'\hat{Q} = Id_K + o_p(1)$ implies that the square matrix $\hat{Q}$ is bounded in probability, and invertible w.p.a. 1. This condition also implies that $[\hat{Q} - (\hat{Q}^{-1})']'[\hat{Q} - (\hat{Q}^{-1})'] = o_p(1)$, i.e. $\hat{Q}^{-1} = \hat{Q}' + o_p(1)$. By pre-multiplying both sides of equation $\hat{Q}'D\hat{Q} = D + o_p(1)$ by $\hat{Q}$, we get:

$$D\hat{Q} = \hat{Q}D + o_p(1). \tag{a.38}$$

Matrix equation (a.38) for element (k, l) becomes $d_k\hat{Q}_{k,l} = \hat{Q}_{k,l}d_l + o_p(1)$. For $k \neq l$, we have $d_k \neq d_l$, and we get $\hat{Q}_{k,l} = o_p(1)$. Then, condition $\hat{Q}'\hat{Q} = Id_K + o_p(1)$ implies $\hat{Q}^2_{k,k} = 1 + o_p(1)$ for any $k = 1, ..., K$. The conclusion follows, with $\hat{S} = diag(sign(\hat{Q}_{k,k}), k = 1, ..., K)$, where $sign(\cdot)$ is the sign function. Q.E.D.