Journal of Econometrics 201 (2017) 176–197

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Double instrumental variable estimation of interaction models with big data

Patrick Gagliardini [a,b,*], Christian Gouriéroux [c,d]

[a] Università della Svizzera Italiana (USI), Lugano, Switzerland
[b] SFI, Switzerland
[c] CREST, France
[d] University of Toronto, Canada

ARTICLE INFO

Article history:
Available online 1 September 2017

JEL classification:
C23
C26
C38
C55

Keywords:
Interaction model
Factor analysis
Big data
Instrumental variable
Recommender system

ABSTRACT

The factor analysis of a (n, m) matrix of observations Y is based on the joint spectral decomposition of the matrix squares YY' and $Y'Y$ for Principal Component Analysis (PCA). For very large matrix dimensions n and m, this approach has a high level of numerical complexity. The big data feature suggests new estimation methods with a smaller degree of numerical complexity. The double Instrumental Variable (IV) approach uses row and column instruments to estimate consistently the factors via an averaging method. We compare the double IV approach to PCA in terms of numerical complexity and statistical efficiency. The double IV approach can be used for the analysis of recommender systems and provides a new collaborative filtering approach.

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 data sets provide detailed information on the interdependencies and interactions between the individual behavior of economic and social agents.

In this paper we consider the interactions in a homogeneous 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 and Worms (2004), Gourieroux et al. (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 and Minehart (2011)], the co-citations between researchers in Economics, the table of import/export to major trading partners [see e.g. Leng and Tang (2012)], and 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 and Khoshgoftaar (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 and Faust (1994), Nowicki and Snijders (2001), Jackson (2008), Iijima and Kamada (2010), Boucher and Mourifie (2013)].

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

[*] Correspondence to: Università della Svizzera Italiana, Via Buffi 13, CH-6900 Lugano, Switzerland.
E-mail address: patrick.gagliardini@usi.ch (P. Gagliardini).

[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 software for PCA and Exploratory Factor Analysis available in SAS.

http://dx.doi.org/10.1016/j.jeconom.2017.08.002
0304-4076/© 2017 Elsevier B.V. All rights reserved.

estimation techniques require a number of 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 interaction models. We consider specifications with unobservable row and column factors. We estimate the factor values by a methodology inspired by the classical Instrumental Variable (IV) approach. We show that this estimation methodology has a smaller degree of numerical complexity compared to PCA, while it achieves the same statistical efficiency when instruments are optimally selected. The present paper may be considered as an introduction of a new model for big data and a new estimation method. The application to real data sets is beyond the scope of this paper and would bring a better balance between explanatory goals vs predictive goals.

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 and Reichlin (1996). Differently from the standard IV framework, instruments can be constructed by partial averaging of nonlinear transformations of the interaction data and do not require exogenous data. We derive the asymptotic properties of these linear IV 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 PCA. 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 with single- and multiple-factor models. 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 $i = 1, \dots, n$, and $j = 1, \dots, m$, respectively. We denote $y_{i,j}$ the magnitude of the interaction from i to j.[3]

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

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

where α_i and β_j are K-dimensional stochastic row and column factors and $\varepsilon_{i,j}$ is a scalar error term. Factor values and error terms are unobservable. The homogeneity assumption is:

Assumption A.1 (*Homogeneity*)**.** Random 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. The independence assumptions are conditional on the knowledge of the number of factors K. Typically, selecting a too small number of factors can induce spurious dependences. In the rest of the theoretical analysis of the paper, we assume that K is the correct number of factors.

For expository purpose, we start by assuming that the factors and errors have zero mean. The extension of the estimation methodology to accommodate non-zero expectations is postponed to Section 2.3.

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 values, and ε the (n, m) matrix of error terms. For a given matrix such as Y, we denote y_i the $(m, 1)$ vector $y_i = (y_{i,j}, j = 1, \dots, m)$, that is the transposed of row i of matrix Y, and by y^j its jth 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 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 and 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 and 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:

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

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 and Nagar (2000), or Leng and Tang (2012)].

The condition of independence between row and column factors α_i and β_j in Assumption A.1 is not essential for our estimation approach. It could be relaxed at the cost of some complications in the asymptotic distribution of the double IV estimator.

[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.

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