Structural Modelling of Dynamic Networks and Identifying Maximum Likelihood

Gourieroux*, C. and J. Jasiak†

January 16, 2023

Abstract

This paper considers nonlinear dynamic models where the main parameter of interest is a nonnegative matrix characterizing the network (contagion) effects. This network matrix is usually constrained either by assuming a limited number of nonzero elements (sparsity), or by considering a reduced rank approach for nonnegative matrix factorization (NMF). We follow the latter approach and develop a new probabilistic NMF method. We introduce a new Identifying Maximum Likelihood (IML) method for consistent estimation of the identified set of admissible NMF's and derive its asymptotic distribution. Moreover, we propose a maximum likelihood estimator of the parameter matrix for a given non-negative rank, derive its asymptotic distribution and the associated efficiency bound.

Keywords: Network, Nonnegative Matrix Factorization, Set Identification, Heterogeneity, Identifying Maximum Likelihood (IML), Alternating Maximum Likelihood (AML).

*University of Toronto, Toulouse School of Economics and CREST, e-mail:*Christian.Gourieroux@ENSAE.fr.*

†York University, e-mail: *jasiakj@yorku.ca.*

We thank X. D'Haultfoeuille, M. Henry, A. Monfort and the participants of the CREST seminar for helpful comments.

1 Introduction

The network effects are commonly represented by a non-negative matrix A of dimension (n, m). There exists a large literature on network models. It can be divided into research strands that differ with respect to the objective of analysis, which is either prediction-oriented or structural, as well as with respect to the assumptions imposed on the non-negative matrix A. These assumptions may concern the 0 and 1 entries of adjacency matrices, or the positive elements of incidence matrices and their "reduced rank". In applications to either contagion, transmission, social interaction, or spillover effects, the elements of each row of matrix A can be constrained to sum up to one. Then, each row of matrix A is interpreted as a conditional probability distribution and that matrix is called a transition, or migration matrix.

The prediction-oriented techniques are applied to image analysis, facial recognition and machine learning. The dimensions n and m of matrix A are very large, and the primary goal of the method is to reduce its complexity. Hence, adjacency matrices with a small number of links, i.e. entries equal to 1 are used to obtain sparse matrices A. Also, incidence matrices are chosen so that non-negative matrix factorization (NMF) with a rather small reduced rank can be applied [Berman, Plemmons (1994)].

The methods applied to network analysis in econometrics are more structural and parametric [see Manski (1993) for the reflection problem, Blume et al. (2011), de Paula (2017) for a survey and the references therein]. The models often represent an equilibrium, such as:

$$Y_t = AY_t + DX_t + u_t,$$

where $Y_t = (y_{1,t}, ..., y_{n,t})$ are the individual observations at time t, and X_t are the explanatory variables. At the equilibrium, a simultaneity arises because the variable Y appears on both sides of the equation. The dimensions n, m of matrix A are much smaller than in the image analysis. Moreover, in order to obtain a simple parametric model, matrix A is often defined as a parametric nonnegative combination $A = \sum_{l=1}^{L} \alpha_l A_l,\ \alpha_l \geq 0,\ l = 1, \ldots, L$, of known network matrices, which is a crucial assumption in this literature [see e.g. Bramoulle et al. (2009), Lee et al. (2010), De Giorgi et al. (2010), Cohen-Cole et al. (2014), Blume et al. (2015)].

The aim of this paper is to fill the gap between the prediction-oriented and the structural methods by introducing the identifying Maximum Likelihood method, a Maximum Likelihood (ML) estimator of the set of admissible NMFs as well as an ML estimator of matrix A, in an extended class of dynamic probabilistic models.

We introduce the class of dynamic parametric models with interaction matrices of reduced rank, and discuss various examples, such as the static models used in the machine learning literature, the dynamic panel models for individual qualitative histories, and the multivariate dynamic Poisson model with contagion used in epidemiology, as special cases of that extended class of models. The NMF of matrix A in this extended class of models may involve not only the directional factors, but also a latent heterogeneity distribution. In general, there exist multiple admissible factorizations of a given true network matrix A_0 with or without zero entries. We describe analytically the identified set for nonnegative ranks $K = 1$, and $K = 2$ of matrix A and show how it can be parametrized in the general case $K \geq 2$. The proposed statistical inference methods rely on the Identifying Maximum Likelihood approach which is introduced to estimate the set of NMF's and to derive the asymptotic distribution of the estimated identified set. In this context, the estimation of a NMF with the most concentrated latent heterogeneity is also considered. Moreover, we propose a maximum likelihood estimator of the nonnegative matrix A, given its nonnegative rank, and we derive the efficiency bound for the rank constrained matrix parameter.

Our approach differs from the methods used in the existing network literature with respect to the following: 1) In our paper, the network is examined in a nonlinear dynamic framework, allowing us to distinguish the short and long run effects. Our approach is applicable to static linear models considered in the existing literature, which arise as special cases; 2) The matrix A is compatible with the existing empirical applications that mainly concern the semi-aggregated level networks with matrices A without zero elements (intricate network) and the diagonal (resp. off-diagonal) elements representing the within (resp. between) segment connections; 3) In our paper, matrix A is assumed to satisfy a factor decomposition into non-negative factorial directions. To solve the partial identification issue, we derive analytically the identified set and show that it can be locally parametrized. In particular, the identified set is not defined from inequality moment restrictions, as it is commonly assumed in micro-econometric partial identification literature [see Chernozhukov et al. (2007). Andrews, Soares (2010), Canay, Shaikh (2017) for the statistical approach to this partial identification]. Instead, the Identifying Maximum Likelihood (IML) method is introduced in this paper, allowing for sharp estimation of the identified set and of the rank constrained matrix A.

The paper is organized as follows. In Section 2, we introduce the class of dynamic parametric models with interaction matrices of reduced rank. Section 3 discusses the identification of a NMF of matrix A. Statistical inference is developed in Section 4. Section 5 presents the link with the nonparametric approaches to partial identification. Section 6 concludes.

The list of additional regularity assumptions for asymptotic statistical inference is given in on-line appendices.

2 Parametric Model with Interaction Matrix

2.1 The model

We consider a set of observations $Y_t, t = 1, \ldots, T$, that can be scalars, vectors, or matrices. We assume that (Y_t) is a stationary Markov process and introduce a parametric model for the conditional distribution of Y_t given its lagged values, with a conditional probability density function (p.d.f.) :

$$l(y_t|\underline{y_{t-1}}) = l(y_t|y_{t-1}; A), \tag{2.1}$$

where A is an unknown non-negative matrix $A \geq 0$ of dimension (n, m), which has non-negative entries.

We make the following assumption :

Assumption A.1 :

i) The parametric model is well-specified, with a true value A_0 of matrix parameter A.

ii) The process (Y_t) is strictly stationary, geometrically ergodic.

It is well-known that a nonnegative matrix A can be factorized as :

$$A = BC', \tag{2.2}$$

where B (resp.C) have dimensions (n, K) [resp. (m, K)] and are nonnegative : $B \geq 0, C \geq 0$. Among the multiple non-negative factorizations available, some correspond to a minimal order K, called the nonnegative rank of matrix A and denoted by $Rk_+(A)$. The nonnegative rank of A is always larger or equal to the rank of A.

A nonnegative matrix factorization (NMF) can be written under different equivalent forms. Let $\beta_k, k = 1, \ldots, K$ (resp. $\gamma_k, k = 1, \ldots, K$) denote the columns of B (resp. C). We have :

$$B = (\beta_1, \ldots, \beta_K), C = (\gamma_1, \ldots, \gamma_K), \tag{2.3}$$

and then :

$$A = \sum_{k=1}^{K} \beta_k \gamma_k' \text{ with } \beta_k \geq 0, \gamma_k \geq 0, \forall k. \tag{2.4}$$

This provides a decomposition of A as the sum of K non-negative matrices of rank 1.

In structural models, we can be interested not only in the true value A_0, but also in a true NMF : $B_0 C_0' = \sum_{k=1}^{K_0} \beta_{0,k} \gamma_{0,k}'$, that generates A_0. We introduce the additional assumptions :

Assumption A.2 :

i) The nonnegative rank $K_0 = Rk_+(A_0)$ is known.

ii) The true matrix A_0 is asymptotically identifiable, i.e. the optimization problem $\max_A E_0 \log l(Y_t|Y_{t-1}; A)$ where the optimization is with respect to the set of nonnegative matrices, has the unique solution $A = A_0$.

iii) The vectors $\beta_{0,k}, \gamma_{0,k}, k = 1, \ldots, K$, have strictly positive entries.

Remark 1 : The analysis of this paper is easily extended to parametric conditional models including also observed exogenous variables X_t, or additional parameters θ, that is to models of the type :

$$l(y_t|\underline{y_{t-1}}, \underline{x_t}) = l(y_t|y_{t-1}, x_t; A_0, \theta_0),$$

in particular to dynamic panel models with covariates.

2.2 Examples

In general the NMF is applied without assuming a probabilistic structure and our objective is to extend to NMF, what has been done for principal component analysis by Tipping, Bishop (1999). The examples below show models that could be introduced for different types of applications as facial recognition, epidemiology, or credit risk. These models have to account for the nonnegativity of the observations Y_t, usually encountered in practice.

2.2.1 Static model

The static models assume that observations $Y_t, t = 1, .., T$ are independent and identically distributed (i.i.d.). These models are commonly used for image analysis [see Lee, Seung (1999) for the first application to learning the parts of objects] under a non-probabilistic approach. In this application, the observations are matrices $Y_t = (Y_{i,j,t})$, where (i, j) denote

the coordinates of a point in picture t and the value $Y_{i,j,t}$ provides the pixel intensities associated to coordinates i, j and picture t. The pixel intensity can be measured on either a discrete, or continuous scale. Then, the considered model:

$$l(y_t|y_{t-1}; A) = \Pi_{i=1}^{n}\Pi_{j=1}^{m} f(y_{i,j,t}; a_{i,j}), \tag{2.5}$$

is based on a family of probability density functions (p.d.f.) $f(y; a)$ with nonnegative argument y and nonnegative scalar parameter a. For the Poisson and exponential p.d.f., the specification in (2.5) simplifies and leads to a generalized linear model (GLIM) [see McCullagh, Nelder (1989) for GLIM and Collins et al. (2002) for its use for factor analysis].

In other applications, network matrices can be observed and analyzed by static models. For instance, we can consider a set of individuals $i = 1, \ldots, L$ and observe at time t the number $y_{i,j,t}$ of messages sent by individual i to individual j. The i.i.d observations y_t can also be matrices containing the investments of bank i in industrial sector j at time t , the gravity matrices summarizing the international trades between countries [Chen et al. (2021), Section 6], or the matrices representing the numbers of stocks of firms head-quartered in city $j, j = 1, \ldots, m$ and selected by mutual fund manager $i, i = 1, \ldots, n$ at time t [Hong, Xu (2014)].

2.2.2 Panel of individual qualitative histories.

Let us consider a panel of L individuals. Each individual is characterized by a qualitative state $i = 1, \ldots, n$, that can be observed at any time t. The qualitative individual histories can be quantified and represented by n-dimensional vectors $(Y_{l,t}, t = 1, \ldots, T)$, $l = 1, \ldots, L$, where $Y_{l,t}$ has entries that sum up to 1.
The dynamic model can be defined by assuming that :

i) the individual histories are independent;

ii) each individual history corresponds to a Markov chain with transition matrix A.

The above assumptions imply that the population of interest is homogeneous. Under these assumptions the individual histories can be aggregated without a loss of information and replaced by the counts of individuals in each state :

$$Y_t = \sum_{l=1}^{L} Y_{l,t}. \tag{2.6}$$

Then the sequence of multivariate counts Y_t is also a Markov process with the conditional

p.d.f. obtained by considering the convoluate of different multinomial distributions :

$$l(y_t|y_{t-1};A) = \Pi_{i=1}^{n} l_i(y_{i,t}|y_{i,t-1};a_i), \tag{2.7}$$

where l_i denotes the p.d.f. of the multinomial distribution $M(y_{i,t-1};a_i)$ and a_i denotes the i^{th}-row of the transition matrix A.

In this example, we have replaced the individuals by homogeneous segments. For example, the corporates can be replaced by industrial sectors, households by age categories, or animals by species [Donnet, Robin (2021) and the references therein]. Then, the diagonal elements of matrix A depict the interactions within a segment and the off-diagonal elements of A represent the interactions between the segments. In general, the diagonal elements a_{ii} are not equal to zero.

2.2.3 Dynamic model for a non-negative random vector

Let us consider a parametric model for a non-negative random vector $l(y;\theta)$, where y and the parameter vector θ have equal dimension n, and both vectors y and θ are nonnegative. Then, a dynamic model for (y_t) can be defined as :

$$l(y_t|y_{t-1};A) = l(y_t;Ay_{t-1}), \tag{2.8}$$

where the contagion matrix A is of dimension (n,n) and nonnegative.

When the parametric model represents the dynamics of n independent Poisson variables, we get a multivariate Poisson autoregressive model with :

$$\begin{aligned} l(y_t|y_{t-1};A) &= \Pi_{i=1}^{n}\left[\frac{1}{y_{it!}}\exp(-a_i y_{t-1})(a_i y_{t-1})^{y_{it}}\right] \\ &= \Pi_{i=1}^{n}\left(\frac{1}{y_{it!}}(a_i y_{t-1})^{y_{it}}\right)\exp(-e'Ay_{t-1}), \end{aligned} \tag{2.9}$$

where a_i is the i^{th} row of matrix A and e is the vector with unitary elements.
This specification differs from an exponential specification:

$$y_{it}|y_{t-1} \sim \mathcal{P}[\exp(a_i y_{t-1})],$$

considered in Chen et al. (2021), Example 3, and Section 6. This alternative specification does not require matrix A to be nonnegative. However, while the dynamic of model (2.9) is compatible with the stationarity of the process (y_t), the above alternative approach of Chen et al. (2021) leads to explosive trajectories due to the exponential transform.

When the parametric model represents the dynamic of n independent exponential variables, we get a multivariate exponential autoregressive model:

$$l(y_t|y_{t-1}; A) = \Pi_{i=1}^{n} \left[(a_i y_{t-1}) \exp(-a_i y_{t-1} y_{it})\right]. \tag{2.10}$$

This dynamic model can be used to study the joint evolutions of the gross domestic product in a set of L countries. Then, matrix A is unobserved and needs be estimated under some mild constraints to provide a proxy of an international trading network.

Remark 2 : The dynamic specification (2.8) can be extended to :

$$l(y_t|y_{t-1}; A) = l(y_t; Az_{t-1}),$$

where z_{t-1} is a nonnegative vector function of y_{t-1}. Such a transformation appears in structural models used in epidemiology and in other applications such as the analysis of cyber-attacks [Fahrenwaldt et al. (2018), Hillairet, Lopez (2020)], adoption of a new technologies [Brock, Durlauf (2010), Blume et al. (2011)] and the Susceptible-Infected-Recovery (SIR) model with multiple transmissions where the components of y are the counts of infected individuals [resp. of the new adoptions of the product] in different segments of the populations. In the SIR model, z is a quadratic function of y [see e.g. Gourieroux, Jasiak (2022)].

Remark 3: Some time series of observed networks can be available. They consists of nonnegative matrices Y_t that could be transformed into a series of nonnegative vectors by considering $y_t = vec(Y_t)$, to which the dynamic model (2.8) could be applied. However, such a dynamic model would likely encounter the curse of dimensionality and the low rank assumption on A, that is the NMF representation, would lose its main structural interpretation. The analysis of network dynamic is an emerging field. The modelling requires a precise description of the network structure.

2.2.4 Contagion of defaults

The structural models of corporate defaults assume that a default arises when the liability of a firm falls below its asset value. When only the defaults are observed, the individual assets and liabilities are latent nonnegative variables which can be represented by a dynamic network model (2.8):

$$\tilde{l}(z_t|z_{t-1}; A) = \tilde{l}(z_t; Az_{t-1}), \tag{2.11}$$

where z_t has dimension $2L$ and L is the number of firms. The elements of z_t represent the assets and liabilities of firms. For example, z_{1lt}, z_{2lt} are the characteristics of firm l at date t. The observed variables are the defaults, represented by vector y_t of dimension L and defined by :

$$y_{lt} = 1, \text{ if } z_{1lt} > z_{2lt}, = 0, \text{ otherwise.} \tag{2.12}$$

Equations (2.11) and (2.12) can be interpreted as a state space model. From the state equation (2.11) and the measurement equation (2.12) it follows that the transition of the observed variables is $l(y_t|\underline{y_{t-1}}; A)$, involving the complete histories of default of all firms.

All dynamic models given above have nonlinear dynamics due to the nonnegativity constraints on the values of observed variables. They are also nonlinear with respect to parameter A. These nonlinear features distinguish the class of models considered in this paper from the major part of literature on networks in econometrics. [3].

The identification in this extended class of models eliminates the use of powers of matrix A. Indeed, matrices A, A^2, A^3 no longer play a special role, as the conditional distribution of y_t given y_{t-h} does not necessarily depend on A through A^h only [see e.g. de Paula (2017) p 272 for a discussion]. In fact all information on A is captured by the lag 1 of variables in the likelihood function.

Moreover, the response of a nonlinear dynamic system to a shock at date t depends on the current environment of y_t. The size of the shock effect depends on the environment, so that even a small shock can have a large impact. This has important consequences concerning the treatment of small values of the elements a_{ij} of matrix A. More precisely, even if a_{ij} is small, this element should not to be set equal to zero artificially, by applying, for example, an automatic LASSO penalty [see e.g de Paula et al. (2020) p. 285-286]. Especially when the dynamic system is close to a tipping point, a small connection can become the source of a significant change in the system, as observed in the histories of corporate defaults, chains of business failures, or inter-bank liquidity shortages.

[3]"Identification and measurement of network phenomena has drawn attention in fields as diverse as macroeconomics, finance not discussed in the review" [De Paula (2017)]. These other fields with nonlinear dynamic models are also left out of the recent survey on partial identification by Kline, Tamer (2022). The nonlinear dynamic models in our paper are introduced for application to these other fields that also include ecological economy, monetary economy and epidemiology.

2.3 Latent Heterogeneity and Ranking

2.3.1 Alternative parametrization

The NMF (2.4) can be normalized and written alternatively as :

$$A = a \sum_{k=1}^{K} \pi_k \beta_k^* \gamma_k^{*'}, \tag{2.13}$$

where $a = e'Ae = \Sigma_i \Sigma_j a_{i,j} > 0, \pi_k \geq 0, k = 1, \ldots, K$, with $\sum_{k=1}^{K} \pi_k = 1, \beta_k^* \geq 0, \gamma_k^* \geq 0, k = 1, \ldots, K$ with : $\beta_k^{*'} e = \gamma_k^{*'} e = 1, k = 1, \ldots, K$.

In this decomposition $\pi = (\pi_1, \ldots, \pi_K)', \beta_k^*, \gamma_k^*, k = 1, \ldots, K$ can be interpreted as discrete probability distributions. More precisely, the normalised matrix A/a can be interpreted as a joint probability distribution, and its decomposition $\sum_{k=1}^{K} \pi_k \beta_k^* \gamma_k^{*'}$ as a mixture of independent joint distributions. In this respect, our analysis is linked to the literature on partial identification of finite mixtures [see e.g. Hall, Zhou (2003), Kasahara, Shimotsu (2009), Henry et al. (2014) and Section 5]. When $K = 1$, the NMF becomes : $A = a\beta_1^* \gamma_1^{*'}$.

The representation (2.13) solves the identification issue due to the identification of factorial directions up to positive scalars. It also allows to bound the set of NMF's written under this form.

2.3.2 Rankings

To motivate the alternative parametrization (2.13), let us consider a dynamic model of count variables for epidemiology. The dynamic contagion model (2.8) can be considered, where the components of y_t are the counts of infected individuals in L homogenous segments of the population. In addition, let us assume that [4] :

$$E_{t-1} y_t = A y_{t-1}. \tag{2.14}$$

If $K = 1$, we get : $E_{t-1} y_t = a\beta_1^* \gamma_1^{*'} y_{t-1}$, or equivalently :

$$E_{t-1} y_{i,t} = \sum_{j=1}^{L} a_{ij} y_{j,t-1} = a\beta_{1i}^* \sum_{j=1}^{L} \gamma_{1j}^* y_{j,t-1}. \tag{2.15}$$

[4] This constraint on the conditional mean implies $y_t = Ay_{t-1} + u_t$, where u_t is a martingale difference sequence with $E_{t-1}(u_t) = 0$. It does not imply a linear dynamic model : $y_t = Ay_{t-1} + u_t$, where u_t is a strong (i.i.d.) white noise. Hence, it is inadequate for the analysis of nonlinear shock effects, i.e. nonlinear impulse response functions.

The contagion parameters a_{ij} can be decomposed as: $a_{ij} = a\beta^*_{1i}\gamma^*_{1j}$, where a is a global contagion effect, β^*_{1i} an index of vulnerability of segment i to the infection and γ^*_{1j} a measure of viral load of segment j. [5] Therefore the segments $i = 1, \ldots, L$ can be ranked with respect to their vulnerability β^*_{1i} and their infectiosity γ^*_{1i}.

When K is larger or equal to 2, a latent heterogeneity of segments arises with heterogeneity distribution π. Then, the segments can be ranked with respect to different notions of vulnerabilities, i.e. the $\beta^*_{ki}, k = 1, \ldots, K$, and infectiosity, i.e. the $\gamma^*_{k,i}, k = 1, \ldots, K$.

The potential interpretations of parameters β^*, γ^*, π depend on the application of interest as shown below.

Example 1 : In the analysis of internet diffusion of messages (see Section 2.2.1), parameters β^* (resp. γ^*) can be used for ranking of receivers and senders, or followers and influencers, in trade networks for ranking the importers and exporters, in citation networks for ranking the citees and citors.

3 Identification of the Nonnegative Factorization

The true NMF is not point-identified. This section discusses the identification issues and derives the identified set for small nonnegative ranks.

3.1 The general framework

The parametric model depends on the nonnegative matrix factorization :

$$A = BC' = (\beta_1, \ldots, \beta_K)(\gamma_1, \ldots, \gamma_K)' = \sum_{k=1}^{K} \beta_k \gamma_k',$$

where K denotes the nonnegative rank and $\beta_k, \gamma_k' s$ are the factorial directions. In practice, the structural parameters are $K, \beta_k, \gamma_k, k = 1, \ldots, K$, and there exists a large body of literature on the lack of identification of these parameters for a given matrix A.

It is easy to see that the factorization is not unique because the same matrix A is obtained from a permutation of index k and rescaling by a positive scalar, i.e. by replacing

[5] The decomposition of a_{ij} is multiplicative, while an additive decomposition: $a_{ij} = \tilde{\alpha} + \tilde{\beta}_i + \tilde{\gamma}_j$, is used in the panel literature with two-ways fixed effects [see e.g. the running examples in Fernandez-Val, Weidner (2016) for probit and Poisson models]. In our framework it is not possible to transform the multiplicative form into an additive one by taking the logarithms of $a_{ij}, \beta^*_{1i}, \gamma^*_{1j}$, as there can exist individuals with zero value of β^*_1, such as the vaccinated or naturally immunized individuals. From the identification perspective, these zero values can be informative and should not be disregarded. Moreover, as noted in Chen et al. (2021), such interactive effects can capture network features as homophily and clustering.

β_k, γ_k by $\sigma_k\beta_k, \gamma_k/\sigma_k$ for a positive scalar σ_k. This identification issue is easily solved by a normalization [a free normalization in the terminology of Lewbel (2019), Section 6.3]. The more complicated identification issues that arise when $K \geq 2$ are discussed below under the following assumption :

Assumption A.3 : The nonnegative rank of matrix A is equal to the rank of A.

This assumption is not very stringent, even though some examples of nonnegative matrices with $Rk_+(A) > Rk(A)$ have been given in the literature. It is useful to describe and parametrize the set of admissible NMF's when $Rk_+(A) = Rk(A)$.

Assumption A.3 can be written under different equivalent forms :

Assumption A.3 is satisfied

$\Longleftrightarrow$ the vectors $\beta_1, \ldots, \beta_K$ are linearly independent and

the vectors $\gamma_1, \ldots, \gamma_K$ are linearly independent

$\Longleftrightarrow$ $B'B$ and $C'C$ are invertible. [6]

Assumption A.3 implies that the matrices $\beta_k\gamma_l', k, l = 1, \ldots, K$ are linearly independent as shown in Lemma 1 below :

Lemma 1 : Under Assumption A.3, $B\Delta C' = 0 \Rightarrow \Delta = 0$.

Proof : Indeed $B\Delta C' = 0$ implies $B'B\Delta C'C = 0$, and then $\Delta = 0$, since $B'B$ and $C'C$ are invertible.

QED

Under assumptions A.1-A.3, the nonnegative rank is known as well as the rank of A_0. Let us first consider another factorization of matrix A without taking into account the nonnegativity conditions on $\beta_k, \gamma_k's$. Since the range of A (resp. A') is the space spanned by $\beta_1, \ldots, \beta_K$ (resp. by $\gamma_1, \ldots, \gamma_K$), an alternative factorization is :

$$A = B\ G\ H'C' = B^*C^{*'},$$

where $B^* = B\ G, C^* = C\ H$ and G, H are invertible matrices [7]. Moreover, we have :

[6] The NMF representation is different from the Singular Value Decomposition (SVD) of matrix A. In SVD the identification issue is usually solved by introducing the orthonormality restriction $B'B = C'C = Id$. Orthogonality is not possible in our framework since $\beta_k'\beta_l$ is always nonnegative. Then the orthogonality condition would imply $\beta_{ik}\beta_{il} = 0, \forall i$ and contradict assumption A.1.

[7] The columns of B and the columns of B^* are two bases of the range of AA' equal to the range of A. Therefore they satisfy a one-to-one relationship represented by an invertible matrix G.

$$B\ G\ H'\ C' = B\ C',$$

and by Lemma 1 we deduce that $H' = G^{-1}$. Therefore we get :

$$B^* = BG, C^* = C(G')^{-1}.$$

In addition, because of the definition of factors up to the permutation and (signed) scale effects, we can choose G of the type :

$$G = Q \text{ diag } \sigma, \tag{3.1}$$

where $\sigma = (\sigma_k), \sigma_k > 0$, and Q a (K, K) matrix with diagonal elements equal to 1. Then, $(G')^{-1} = (Q')^{-1}$ diag $(1/\sigma)$.

Then, by taking into account the nonnegativity conditions, we get a constructive characterization of the identified set :

Proposition 1 : For a specific factorization of matrix A_0 : $A_0 = B_0C_0', B_0 \geq 0, C_0 \geq 0$ (referred to as a specific element of the identified set), all observationally equivalent nonnegative factorizations are such that :

$$B^* = B_0Q \text{ diag } \sigma,\ C^* = C_0(Q')^{-1} \text{ diag } (1/\sigma),\ \sigma_k > 0, \forall k = 1, \ldots, K,$$

where the matrix Q is invertible, with unitary diagonal elements and such that :

$$B_0Q \geq\ 0,\ C_0(Q')^{-1} \geq 0.$$

The nonnegative factorization is said to be essentially unique, or simply unique [see Laurberg et al. (2008)], if $Q = Id$ is the only solution to the set of inequalities given in Proposition 1.

3.2 Special Cases

3.2.1 Case $K = Rk_+(A) = 1$

When $K = 1, A = \beta_1\gamma_1', Q = (q_{11}) = (1)$, the nonnegative factorization is essentially unique, i.e. the NMF is (essentially) point identified. Moreover, we note that :

$$A\beta_1 = \beta_1(\gamma_1'\beta_1), A'\gamma_1 = \gamma_1(\beta_1'\gamma_1).$$

It follows that β_1 (resp. γ_1) is an eigenvector of A (resp. A') associated with the eigenvalue $\gamma_1'\beta_1 = \beta_1'\gamma_1$, which is strictly positive. By the Perron-Froebenius Theorem [see Meyer (2000)] the nonnegative matrix A (resp. A') has a unique eigenspace of dimension 1 generated by a nonnegative eigenvector, here β_1 (resp. γ_1) [8].

3.2.2 Case $K = Rk_+(A) = 2$

For $K = 2$, we get $Q = \begin{pmatrix} 1 & q_{12} \\ q_{21} & 1 \end{pmatrix}$, and

$$(Q')^{-1} = \frac{1}{1 - q_{12}q_{21}} \begin{pmatrix} 1 & -q_{21} \\ -q_{12} & 1 \end{pmatrix}.$$

Without the subscript 0 for the "specific factorization of true matrix", the inequality restrictions in Proposition 1 become :

$$\begin{cases} \beta_{1,j} + q_{21}\beta_{2,j} \geq 0,\ j = 1, \ldots, n,\ q_{12}\beta_{1,j} + \beta_{2,j} \geq 0,\ j = 1, \ldots, n, \\ \\ \dfrac{1}{1 - q_{12}q_{21}}(\gamma_{1,j} - q_{12}\gamma_{2,j}) \geq 0, j = 1, \ldots, m, \dfrac{1}{1 - q_{12}q_{12}}(-q_{21}\gamma_{1,j} + \gamma_{2,j}) \geq 0, \\ \\ j = 1, \ldots, m. \end{cases}$$

It is easy to check that these inequalities imply $1 - q_{12}q_{21} > 0$. This inequality is in particular satisfied when we focus our attention on the local identification in a neighbourhood of $(\beta_1, \beta_2), (\gamma_1, \gamma_2)$, i.e. in a neighbourhood of $q_{12} = q_{21} = 0$. Then, the system of inequalities is equivalent to :

$$\begin{cases} \beta_{1,j} + q_{21}\beta_{2,j} \geq 0, \forall j, \text{ with } \beta_{2,j} > 0,\ -q_{21}\gamma_{1,j} + \gamma_{2,j} \geq 0, \forall j, \text{ with } \gamma_{1,j} > 0, \\ \\ q_{12}\beta_{1,j} + \beta_{2,j} \geq 0,\ \forall j, \text{ with } \beta_{1,j} > 0,\ \gamma_{1,j} - q_{12}\gamma_{2,j} \geq 0, \forall j, \text{ with } \gamma_{2,j} > 0. \end{cases}$$

or

$$\begin{cases} q_{21} \geq \sup_{j:\beta_{2,j}>0}(-\beta_{1,j}/\beta_{2,j}),\ q_{12} \leq \inf_{j:\gamma_{1,j}>0}(\gamma_{2,j}/\gamma_{1,j}), \\ \\ q_{12} \geq \sup_{j:\beta_{1,j}>0}(-\beta_{2,j}/\beta_{1,j}),\ q_{21} \leq \inf_{j:\gamma_{2,j}>0}(\gamma_{1,j}/\gamma_{2,j}). \end{cases}$$

We deduce the following result :

[8] It is easy to check that β_1 (resp. γ_1) is also an eigenvector of AA'(resp. $A'A$). Therefore they are elements of a singular value decomposition (SVD) (see Remark 3 in Section 4.1). Such a model is largely applied in the network literature with adjacency matrix, where the leading left and right eigenvectors of the SVD are used to define the "so-called" hub and authority centralities, respectively [see Cai et al. (2022) and the references therein].

Proposition 2 : For $K = Rk_+(A) = 2$, the admissible matrices $Q = \begin{pmatrix} 1 & q_{12} \\ q_{21} & 1 \end{pmatrix}$ are such that :

$$-\inf_{j:\beta_{2,j}>0}(\beta_{1,j}/\beta_{2,j}) \leq q_{21} \leq \inf_{j:\gamma_{1,j}>0}(\gamma_{2,j}/\gamma_{1,j}),$$

$$-\inf_{j:\beta_{1,j}>0}(\beta_{2,j}/\beta_{1,j}) \leq q_{12} \leq \inf_{j:\gamma_{2,j}>0}(\gamma_{1,j}/\gamma_{2,j}).$$

Therefore, among the $2(n+m)$ inequality restrictions in Proposition 1, only four are active and the remaining ones are redundant.

We deduce the necessary and sufficient exclusion conditions for essential uniqueness [see e.g. Brie (2015)].

Corollary 1 : Under Assumption A.3 and for $K = Rk_+(A) = 2$, the nonnegative factorization is essentially unique if and only if there exists at least one index j_1 such that $\beta_{1,j_1} = 0, \beta_{2,j_1} > 0$, one index j_2 such that $\beta_{1,j_2} > 0, \beta_{2,j_2} = 0$, one index j_3 such that $\gamma_{1,j_3} = 0, \gamma_{2,j_3} > 0$ and one index j_4 such that $\gamma_{1,j_4} > 0, \gamma_{2,j_4} = 0$.

Without the "exclusion" restrictions of Corollary 1 and in particular under Assumption A.2 iii), there exists a multiplicity of admissible nonnegative factorizations that are easily deduced from one of them. This is the case we are interested in.

Let us now discuss the degree of underidentification. Since the identification issue due to the product of factorial directions by nonnegative scalars is solved in representation (2.13), we can focus on the identification of $\pi_k, \beta_k^*, \gamma_k^*, k = 1, 2$. For K=2, the decomposition (2.13) after transformation a is:

$$A = a[\tilde{\pi}_1 \tilde{\beta}_1^* \tilde{\gamma}_1^{*'} + \tilde{\pi}_2 \tilde{\beta}_2^* \tilde{\gamma}_2^{*'}].$$

It is easy to check (see online Appendix 1) that :

$$\tilde{\beta}_1^* = p_1 \beta_1^* + (1-p_1)\beta_2^*, \text{ with } p_1 = \beta_1' e/(\beta_1' e + q_{21}\beta_2' e),$$

$$\tilde{\beta}_2^* = p_2 \beta_1^* + (1-p_2)\beta_2^*, \text{ with } p_2 = q_{12}\beta_1' e/(q_{12}\beta_1' e + \beta_2' e),$$

$$\tilde{\gamma}_1^* = p_3 \gamma_1^* + (1-p_3)\gamma_2^*, \text{ with } p_3 = \gamma_1' e/(\gamma_1' e - q_{12}\gamma_2' e),$$

$$\tilde{\gamma}_2^* = p_4 \gamma_1^* + (1-p_4)\gamma_2^*, \text{ with } p_4 = -q_{21}\gamma_1' e/(q_{21}\gamma_1' e - \gamma_2' e),$$

and

$$\tilde{\pi}_1/\tilde{\pi}_2 = (\beta_1' e + q_{21}\beta_2' e)(\gamma_1' e - q_{12}\gamma_2' e)/(q_{12}\beta_1' e + \beta_2' e)(-q_{21}\gamma_1' e + \gamma_2' e).$$

Corollary 2 : For $Rk(A) = Rk_+(A) = 2$, the components $\pi_k, \beta_k^*, \gamma_k^*, k = 1, 2$ are not identifiable, with a degree of underidentification equal to 2.

The set of admissible decompositions (2.13) is described by means of q_{12}, q_{21}. The conditions on q_{12}, q_{21} derived in Proposition 2 remain valid when β, γ are replaced by β^*, γ^*.

3.2.3 General Case

For $K = 2$, the identified set is described by two parameters q_{12}, q_{21} satisfying $2(m + n)$ inequality restrictions. These restrictions are linear and Proposition 2 shows that only 4 of them are active.

In the general case, the identified set is parametrized by $K(K-1)$ parameters that are the off-diagonal elements of matrix Q. Therefore it defines a manifold of fixed dimension. These parameters satisfy $K(n + m)$ inequality restrictions by Proposition 1. The degree of underidentification increases rather quickly with the nonnegative rank and the subset of active restrictions cannot be derived analytically. To determine these restrictions providing a simplified definition of the identified set, numerical algorithms are needed such as Active Set Sequential Quadratic Programming algorithms, some of them being available from Artelys Knitro.

A similar problem arises in sharp set identification for discrete choice models and treatment effects analysis. The main difference is [9] that those inequality restrictions are linear as for $K = 2$ in Section 3.2.2.[10], while in our framework they are nonlinear when $K \geq 3$. Indeed, for $K = 3$, we get : $Q = \begin{pmatrix} 1 & q_{12} & q_{13} \\ q_{21} & 1 & q_{23} \\ q_{31} & q_{32} & 1 \end{pmatrix}$ and, up to the determinant, matrix $(Q')^{-1}$ has cofactor elements, such as $1 - q_{23}q_{32}$ for instance, that are quadratic in Q (more generally, they are polynomials of degree less or equal to $K - 1$). Moreover, it is easy to see that the identified set is not convex.

4 Statistical Inference

A major challenge for statistical inference is the lack of identification of the true NMF. This identification issue can be addressed either by introducing identification restrictions and applying the standard maximum likelihood approach, or by estimating the set of all identifiable NMF directly. These two approaches are linked. For example, the set of all

[9] Another difference is the parametric assumption being used in our framework and nonparametric methods being used in the existing literature (see Section 5).

[10] See, the linear moment functions defining the inequalities in Chernozhukov et al. (2007), Section 2.1.

identifiable NMF's can be deduced from one of them, as shown in Section 3.3.2 for $Rk_+(A) = 2$. Therefore, we can first identify one NMF and next deduce all the remaining ones from the identified one.

In this respect the identified set described in Proposition 1 has the general form [11] considered in Shi, Shum (2015) eq. (1.1)-(1.2). However, their results can only be used if this specific NMF is well-defined and has a consistent and asymptotically normally distributed estimator. These are the points considered in this section.

The proposed method proceeds as follows: 1) We show the convergence of the set of maximum likelihood estimators of B, C to the identified set; 2) For a fixed number of observations, a multiplicity of ML estimators is obtained. Therefore, we introduce an alternating ML algorithm to fix the selected ML estimator for any T. This sequence of alternating ML estimators is well-defined, it converges to the identified set, but not pointwise to a given element of this set; 3) Next, we fix a benchmark in the identified set, which is the solution of an auxiliary optimization. That objective function is optimized with respect to an alternative parametrization in $(a, \pi, \beta^*, \gamma^*)$. This step provides a consistent estimator of a specific element of the identified set.

The proposed method, called the identifying maximum likelihood (IML) approach is used to estimate the identified set. We derive its asymptotic distributional properties and show how to estimate the identified set and analyze its key properties. In addition, we derive the ML estimator of matrix A under a given nonegative rank constraint and the associated efficiency bound.

The approach is feasible because of the properties of the maximum likelihood estimator reviewed below.

4.1 The ML approach

4.1.1 Consistency

Let us assume that the network model is well-specified and the true transition is :

$$l(y_t|y_{t-1}; A_0) = l(y_t|y_{t-1}; B_0 C_0'), \tag{4.1}$$

with a nonnegative rank K_0 of matrix A_0 assumed to be known. We consider a constrained

[11] up to the introduction of nuisance slackness parameters [Shi, Shum (2015), Remark p. 497].

ML maximization, providing:

$$(\hat{B}_T, \hat{C}_T) = \arg \max_{B \geq 0, C \geq 0} \sum_{t=1}^{T} \log l(y_t | y_{t-1}; BC'). \tag{4.2}$$

Due to the identification issue, there exists a large multiplicity of solutions to the finite sample optimization (4.2). Under additional standard regularity conditions given in online Appendix 2, we can derive a consistency property of the above set of ML estimators.

Proposition 3 : Under Assumptions A.1-A.3 and a.1 in Appendix 2, the set of ML estimators $(\hat{B}_T, \hat{C}_T)$ converges to the set $\mathcal{A}_0$ of NMF associated with $A_0 = B_0 C_0'$, when T tends to infinity.

More precisely, let $d(.,.)$ denote the Euclidean distance on $I\!R^{K(n+m)}$, $NMF_0 = \{(B, C), B \geq 0, C \geq 0$, with $BC' = B_0 C_0'\}$, and $\mathcal{D}[(\hat{B}_T, \hat{C}_T), NMF_0] = \min_{(B,C) \in NMF_0} d[(\hat{B}_T, \hat{C}_T), (B, C)]$, then $\mathcal{D}[(\hat{B}_T, \hat{C}_T), NMF_0]$ tends to zero, when T tends to infinity. Under the regularity conditions, this convergence is uniform in $(\hat{B}_T, \hat{C}_T)$. It will also imply the convergence to this set of the well-defined alternating ML estimator introduced in the next section.

Thus, the $(\hat{B}_T, \hat{C}_T)$ does not necessarily converge to the true factorization B_0, C_0 due to the identification issue, but for a large $T, (\hat{B}_T, \hat{C}_T)$ is close to another admissible NMF that can depend on T.

4.1.2 Alternating ML (AML) Algorithm

A ML estimator does not always have a closed-form. Hence, in practice it is computed numerically from an algorithm, such as a Newton-Raphson type of algorithm. In our framework of partial identification, a Newton-Raphson type of algorithm cannot be used jointly for B and C. The reason is that each iteration requires the inversion of a Hessian matrix that is not invertible due to the identification issue. The AML algorithm solves this issue.

In the presence of a multiplicity of NMFs, we apply an alternating AML algorithm [Gourieroux, Monfort, Renault (1990), Kim, Park (2007)]. We observe that, even if the factorization B, C is not identifiable, B (resp. C) is identifiable when C is known (resp. B is known). This leads to the following alternating AML algorithm, where at step p, $\hat{B}_{p,T}, \hat{C}_{pT}$ is computed and then $\hat{B}_{p+1,T}, \hat{C}_{p+1,T}$ are obtained as follows:

$$\hat{B}_{p+1,T} = \arg\max_{B\geq 0} \sum_{t=1}^{T} \log l(y_t|y_{t-1}; B\hat{C}'_{p,T}), \tag{4.3}$$

$$\hat{C}_{p+1,T} = \arg\max_{C\geq 0} \sum_{t=1}^{T} \log l(y_t|y_{t-1}; \hat{B}_{p+1,T}C'). \tag{4.4}$$

By construction, this algorithm produces at each iteration p a higher value of the log-likelihood function than at iteration $p-1$.

We have to distinguish the ML estimator from the alternating ML estimator obtained from the algorithm (4.3)-(4.4). As mentioned earlier, when some parameters are not identifiable there is a multiplicity of ML estimators. However, there is a unique sequence of alternating AML estimators for the given starting values, even though the AML algorithm does not necessarily numerically converge pointwise, due to the identification issue.

4.2 Nonnegative rank $K_0 = Rk_+(A_0) = 1$

As pointed out in Section 3.2.1, the NMF is essentially unique for $K_0 = 1$. This assumption $K_0 = 1$ greatly simplifies the estimation and explains its use in applied econometrics [see e.g. Cai et al. (2022)]. Because the ML estimator is unique in this case, it can be computed from a standard Newton-Raphson algorithm.

We introduce the identification restrictions: $A = a\beta\gamma'$, with $a > 0, \beta \geq 0, \gamma \geq 0$ and $\beta' e = \gamma' e = 1$. Let us now consider the dynamic model described in Section 2.2.3. and based on the latent parametric model $l(y;\theta)$, where θ is replaced by Ay_{t-1} [see, Section 2.2.3]. We get :

$$l(y_t|y_{t-1}; A) = l(y_t; a\beta\gamma' y_{t-1}). \tag{4.5}$$

The partial derivatives of the log-likelihood with respect to a, β, γ are easily derived from the partial derivatives of the latent log-likelihood with respect to θ. We have :

$$\begin{cases} \dfrac{\partial \log l}{\partial \beta}(y_t|y_{t-1}; A) = a\gamma' y_{t-1} \dfrac{\partial \log l}{\partial \theta}(y_t; a\beta\gamma' y_{t-1}), \\ \dfrac{\partial \log l}{\partial \gamma}(y_t|y_{t-1}; A) = a y_{t-1}\beta' \dfrac{\partial \log l}{\partial \theta}(y_t; a\beta\gamma' y_{t-1}), \\ \dfrac{\partial \log l}{\partial a}(y_t|y_{t-1}; A) = \gamma' y_{t-1}\beta' \dfrac{\partial \log l}{\partial \theta}(y_t; a\beta\gamma' y_{t-1}). \end{cases} \tag{4.6}$$

The asymptotic properties, especially the asymptotic distribution of the ML estimators of a, β, γ, depend on the location of the true vectors β_0, γ_0. These properties are straightforward under the assumption A.2 iii) below.

Positivity Assumption A.2 iii): The entries of β_0 and γ_0 are strictly positive.

Under this positivity assumption, the non-negativity-constrained ML estimators have asymptotically strictly positive entries, and the unconstrained and non-negativity constrained estimators are asymptotically equivalent. However, the ML estimator has to account for the linear constraint of unit sum. This estimator without the non-negativity restrictions is defined as :

$$\begin{aligned}(\hat{a}, \hat{\beta}, \hat{\gamma}) &= \arg\max_{a,\beta,\gamma} \sum_{t=1}^{T} \log l(y_t; a\beta\gamma' y_{t-1}), \\ \text{s.t} : \beta' e &= \gamma' e = 1.\end{aligned}$$

The first-order conditions for the Lagrange multipliers associated with the linear restrictions and denoted by λ, μ , are :

$$\sum_{t=1}^{T} [a\gamma' y_{t-1} \frac{\partial \log l}{\partial \theta}(y_t; a\beta\gamma' y_{t-1})] - \lambda = 0,$$

$$\sum_{t=1}^{T} [a y_{t-1}\beta' \frac{\partial \log l}{\partial \theta}(y_t; a\beta\gamma' y_{t-1})] - \mu = 0,$$

$$\sum_{t=1}^{T} [\gamma' y_{t-1}\beta' \frac{\partial \log l}{\partial \theta}(y_t; a\beta\gamma' y_{t-1})] = 0,$$

$$\beta' e = \gamma' e = 1.$$

The FOC need to be solved in $a, \beta, \gamma, \lambda, \mu$.

Asymptotically, when T tends to infinity, we get consistent and asymptotically normal ML estimators of B and C. Their asymptotic variance-covariance matrix has a standard form [see Gourieroux, Monfort (1995), Section 10.3] and its estimate can be computed by standard software.

Remark 4: As pointed out in Section 3.2.1, β_1 (resp. γ_1) can be interpreted as an eigenvector of A (resp. A'). It is easy to check that β_1 (resp. γ_1) is an eigenvector of AA'

(resp. $A'A$). This implies that $A = \beta_1 \gamma_1'$ is a singular value decomposition (SVD) of matrix A, for which statistical inference is available mainly in a Gaussian framework [Anderson, Rubin (1956), Anderson (1963), Tipping, Bishop (1999)].

However, for $Rk(A_0) = Rk_+(A_0) = 1$, the SVD estimation method is not relevant, as its standard asymptotic properties do not account for the non-negativity of the data and the non-negativity of matrix A_0. Moreover, the SVD interpretation of the NMF is no longer valid for $Rk_+(A_0) \geq 2$. Indeed, matrix AA' (resp. $A'A$) is also non-negative, and, except the Perron, Froebenius eigenvector β_1 of AA', all other eigenvectors β_2, β_3 must have at least one negative, or non-real component.

4.3 Nonnegative rank $K_0 = Rk_+(A_0) \geq 2$.

As mentioned in Section 3.2.3, it is sufficient to estimate one of the admissible NMF's to deduce the sharp identified set. The problem with applying the AML method is that the convergence of the AML estimator to the identifed set does not imply its convergence to a given NMF.

This section shows how an IML algorithm can solve this issue, allowing us to derive the asymptotic distribution of the set of admissible NMF's. For expository purpose, we provide in the text the assumptions specific to our problem. The additional assumptions needed for asymptotic analysis are given in online Appendix 2.

4.3.1 Consistency of the alternating ML estimator

In this section we consider the consistency of the AML approximation to the identified set when T tends to infinity and the number of iterations p_T in the AML algorithm depends on T in a suitable manner.

Let us consider the dynamic model :

$$l(y_t|y_{t-1}; A) = l(y_t; a \sum_{k=1}^{K} \pi_k \beta_k^* \gamma_k^{*'} y_{t-1}), \tag{4.7}$$

with $\beta_k^{*'} e = \gamma_k^{*'} e = 1, k = 1, \ldots, K,\ \pi' e = 1$ and the identified set:

$$\mathcal{A}_0 = \{a, \pi_k, \beta_k^*, \gamma_k^*, k = 1, \ldots, K, \text{ such that } a \sum_{k=1}^{K} \pi_k \beta_k^* \gamma_k^{*'} = A_o\}.$$

where A_0 is the true value of A.

Section 2.2 shows that the underlying parametric families $l(y;\theta), \theta \geq 0$, are often constructed from the products of Poisson, or exponential distributions. Therefore, they satisfy the following assumption :

Assumption A.4 : The underlying log-likelihood $\log l(y;\theta)$ is concave in $\theta, \theta \geq 0$.

Under Assumption A.4 and for any value of the number of observations T , each step of the AML algorithm outlined in Section 4.1.2 leads to a unique solution in (B, C), because the objective function is log-concave in B (resp. C) for a given C (resp. B), and also under the alternative parametrization $(\pi_k, \beta_k^*, \gamma_k^*, k = 1, \ldots, K)$. Therefore, the AML estimator is a function of the underlying (normalized) log-likelihood $\frac{1}{T}L_T(.) = \frac{1}{T}\sum_{t=1}^{T} \log l(y_t;.)$, and of the initial values used in the algorithm (and of the number of iterations p). Let the set of parameters be denoted by $\alpha = (\pi_k, \beta_k^*, \gamma_k^*, k = 1, \ldots, K)$ and the selected initial value by α^o. The AML estimator at iteration p can be written as :

$$\hat{\alpha}_T(\alpha^o, p) = m(\frac{1}{T}L_T(.); \alpha^o, p), \tag{4.8}$$

where m is a deterministic function. Then, for large T, the AML estimator will converge asymptotically to the value :

$$\alpha_\infty(\alpha^o, p) = m[E_0 \log l(Y_t; A_0); \alpha^o, p], \tag{4.9}$$

that belongs in the set $\mathcal{A}_0$ if p is large. Note, that this limiting value can depend on the initial value α^0 used in the algorithm.

More precisely, under Assumptions A.1-A.4 and the additional regularity conditions a.1 given in online Appendix 2, we have the following proposition that is a direct consequence of the numerical consistency of the AML approximation (for $p \to \infty$ and T fixed) and of the uniform convergence in Proposition 3 :

Proposition 4: For large T, there exist a function $c(.)$ and a number of iterations p_T such that, for any $p \geq p_T$:

$$\mathcal{D}[\alpha_T(\alpha^o, p), \mathcal{A}_o] < c(\alpha^o)/T,$$

where $\mathcal{D}$ measures the distance between $\alpha_T(\alpha^o, p)$ and the set $\mathcal{A}_o$.

In practice, we can apply the AML algorithm with a given starting value α^o and a number of

iterations p. The number p needs to be set sufficiently large for Proposition 4 to be satisfied. Then, the asymptotic bias of the alternating ML estimator will be sufficiently small to be negligible in the asymptotic distribution of the IML estimator derived later in the text. Moreover, two optimizations are performed at each step of the algorithm, with respect to B and C, respectively. A Newton-Raphson type of algorithm can be used in each of these optimizations. Under the log-concavity assumption A.4, the Newton-Ralphson algorithm is a special case of the steepest ascent algorithm, which increases the objective function at each step. As the increase of the objective function at each step of the algorithm is ensured, the IML with a fixed number of iterations for the intermediate optimization will also increase the objective function, that is sufficient for the numerical consistency of the AML algorithm under Assumption A.4.

It follows from Proposition 1 that all other elements of $\mathcal{A}_0$ are functions of $\alpha_\infty(\alpha^o; p)$ and the elements of a matrix Q with unitary diagonal elements are such that :

$$B[\alpha_\infty(\alpha^o, p)]Q \geq 0, \;\; C[\alpha_\infty(\alpha^o, p)][Q']^{-1} \geq 0.$$

This defines a set $\mathcal{Q}[\alpha_\infty(\alpha^o, p)]$ of admissible values of transformation Q.

Equivalently, we have a parametric representation of the identified set $\mathcal{A}_0$ of NMF's :

$$\mathcal{A}_0 = \{\alpha : \alpha = \xi[\alpha_\infty(\alpha^o, p), Q], Q \in \mathcal{Q}[\alpha_\infty(\alpha^o, p)]\}, \tag{4.10}$$

where ξ is a known function.

Then, the set $\mathcal{A}_0$ is consistently estimated as:

$$\hat{\mathcal{A}}_T = \{\alpha : \alpha = \xi[\hat{\alpha}_T(\alpha^o, p), Q], Q \in \mathcal{Q}[\hat{\alpha}_T(\alpha^o, p)] \equiv \hat{Q}_T(\alpha^o, p)\}. \tag{4.11}$$

The estimation method presented above will allow us to approximate a possibly p-dependent element of the identified set for a sufficiently large p .

Proposition 4 implies that the above AML estimator is a maximizer of the log-likelihood function. This AML estimator can be used to derive a Monte Carlo confidence set for the identified set based on a quasi-likelihood ratio with the quantiles computed by simulations [see, Chen, Christensen, Tamer (2018), Remark 1, p. 1972]. In our framework, the chosen quasi-prior distribution would need to ensure the existence of the log-likelihood function. For example, it has to satisfy the non-negativity restrictions on matrix A in a Poisson autoregressive model. Moreover, the generic form of the estimated confidence set, i.e. $\{A : L_T(A) \geq \xi\}$,

where L_T denotes the log-likelihood and ξ is the estimated quantile, is hard to determine unless the appropriate parametrization given in Proposition 1 is used.

Our objective is to find a consistent ML estimator of the identified set, derive the asymptotic distribution of the estimated identified set and an asymptotic confidence set for the identified set. We will also find the rank-constrained ML estimator of A that maximizes the log-likelihood function under the constraint $A = BC'$, and the associated asymptotic efficiency bound.

Because of the lack of identification, we do not have the numerical stability of $\hat{\alpha}_T(\alpha_0, p)$, for large p. Therefore, we cannot expect to prove any asymptotic normality of this AML estimator. In order to stabilize the algorithm, we include an additional optimization step.

4.3.2 How to introduce an identification restriction by Identifying Maximum Likelihood

An identification issue is commonly solved either by introducing implicit identification restrictions, or by reparametrizing the model and dividing the parameters into the set of identifiable and non identifiable parameters in an appropriate way (this is the "global" reduced form reparametrization considered in Shi, Shum (2015) and Chen et al. (2018), Section 5.1.1). These approaches are not suitable for our framework, where the parametrization of the identified set depends on a selected element of $\mathcal{A}_0$. Hence, we introduce indirectly $K(K-1)$ identification restrictions.

Let us consider an alternative parametrization method. When a specific element of the identified set is known, the identifiable set is parametrized by $vec^*Q = q$, where vec^*Q denote the stacked elements of Q except for the diagonal elements equal to 1. This parametrization cannot be used, as long as that specific element is unknown. However, when a specific element $\tilde{\alpha}$ say, is given, it defines a new origin and then a new parametrization by q of the identifiable set is obtained.

Let us now determine a benchmark α_0^*. We define a benchmark element of the identified set as the optimizer of an additional criterion with respect to the additional parameter q. Two criteria $\tilde{g}(q, \tilde{\alpha}) = g(\alpha)$, where $\alpha \in \mathcal{A}_0$ arise naturally :

i) The concentration of a discrete probability distribution is usually measured by $\sum_{k=1}^{K}(\pi_k \log \pi_k)$. This quantity is negative, it is minimized for $\pi_k = 1/K, \forall k$, that is a uniform distribution, and it increases to zero with the concentration of the distribution. Therefore, we can choose as the benchmark specific element, the factorization providing the most concentrated latent

heterogeneity defined by :

$$\alpha_0^* = \arg\max_{\alpha \in \mathcal{A}_0} \sum_{k=1}^{K} (\pi_k \log \pi_k).$$

ii) An alternative criterion is :

$$g(\alpha) = \det(\tilde{B}'\tilde{B}), \text{ where } \tilde{B} = (\beta_1^*, \ldots, \beta_K^*),$$

or $g(\alpha) = \det(\tilde{C}'\tilde{C})$, where $\tilde{C} = (\gamma_1^*, \ldots, \gamma_K^*)$.

The above criterion measures the volume of the parallelepiped generated by the columns of $\tilde{B}$ (resp. $\tilde{C}$) [see e.g. Barth (1999)]. The larger this volume, the less "colinear" the columns of $\tilde{B}$ (resp. $\tilde{C}$). [12] These criteria could be used jointly.

To proceed with the algorithm-based identification of the benchmark α_0^* by means of this additional optimization, an additional step needs to be included in the AML algorithm.[13] Let the recursive system in this algorithm be denoted by:

$$\alpha^{(p+1)} = H(\alpha^{(p)}),$$

where H depends on the observations.

Then, the identifying maximum likelihood (IML) algorithm is the following :

step 1: Select an initial value $\alpha^{(0)}$.

step p **:** At step p, a value $\alpha^{(p)}$ is available.

i) Apply the AML algorithm to get a value $\tilde{\alpha}^{(p+1)} = H(\alpha^{(p)})$.

This value is considered as an approximation of a point in the identifiable set.

It can be used to parametrize the set $\hat{\mathcal{A}}^{(p)}$ by q.

ii) Perform the optimisation of the additional criterion to get :

$$q^{(p+1)} = \text{Opt}_{q \in Q^{(p+1)}} \tilde{g}(q; \tilde{\alpha}^{(p+1)}),$$

where $Q^{(p+1)}$ is the domain defined by the inequality restrictions applied with $\tilde{\alpha}^{(p+1)}$. Then the solution is a function of $\tilde{\alpha}^{(p+1)}$, that is : $q^{(p+1)} = q(\tilde{\alpha}^{(p+1)})$, say, where $q(.)$ does not depend on the observations.

[12] These criteria are the analogues of the identification restrictions introduced for SVD: $B'B = C'C = Id$, where all factorial directions are orthonormal.

[13] This additional step is the analogue of step 2 in the estimation approach introduced in Davezies et al. (2022), Section 3.1., where it is applied to an approximate identified set instead of the identified set itself.

iii) Find $\alpha^{(p+1)}$ by transforming $\tilde{\alpha}^{(p+1)}$ with a linear transformation $Q^{(p+1)}$ associated with $q^{(p+1)}$, including this choice of decreasing ordering of index to get $\pi_k^{(p+1)}$. More precisely:

compute $Q^{(p+1)}$ such that $q(p+1) = vec^* Q^{(p+1)}$,

compute $\tilde{B}^{(p+1)}, \tilde{C}^{(p+1)}$ from $\tilde{\alpha}^{(p+1)}$

compute $B^{(p+1)} = \tilde{B}^{(p+1)} Q^{(p+1)}$, $C^{(p+1)} = \tilde{C}^{(p+1)} [Q^{(p+1)'}]^{-1}$

compute $\alpha^{(p+1)}$ from $B^{(p+1)}, C^{(p+1)}$, etc.

The main difference between the AML and IML algorithms concerns the consistency. The AML converges to the set $\mathcal{A}_0$, but the convergence is not pointwise. The IML converges to the given α_0^* that allows us to perform a Taylor expansion of the first-order conditions in order to derive the asymptotic normality as in Shi, Shum (2015).

The IML method requires the availability of algorithms for the optimization of nonlinear functions under a large number of nonlinear inequality restrictions. The recent developments in Active Set Sequential Quadratic Programming (SQP) have largely solved this problem [see e.g. Gill et al. (2002), Nocedal, Wright (2006), and Liu (2005) for a proof of numerical convergence].

Remark 5: The additional intermediate optimization in the IML algorithm does not necessarily have to be introduced starting from the first iteration $p = 1$. It can be introduced later, when p is sufficiently large to get a value close to the identified set, by Proposition 4. In this respect, the IML is used to stabilize pointwise the values obtained from a standard AML algorithm.

4.3.3 Asymptotic distributions

The benchmark α_0^* can be either in the interior of the identifiable set, or on its boundary. The asymptotic normality cannot be expected in the latter case, but standard asymptotic arguments can be used to derive the asymptotic normality of the IML estimator adjusted to reach α_0^*, if α_0^* is in the interior of $\mathcal{A}_0$ and the associated π_{0k}^* are all distinct [14].

To clarify the role of the intermediate optimization in the IML algorithm, let us first consider the standard information matrices. Two matrices appear naturally[15] :

i) The unconstrained information based on A, is:

[14] This additional condition is analogous to the condition of distinct eigenvalues in the joint spectral decomposition of AA' and $A'A$ in the standard SVD.

[15] For expository purpose, we keep the same notation l for the conditional likelihood as a function of A, or a function of α.

$$E_0\left[-\frac{\partial^2 \log l(y_t|y_{t-1};A)}{\partial \text{ vec } A\, \partial \text{ vec } A'}\right].$$

This matrix is invertible by the assumption of identifiable A, but can be of a high dimension in practice.

ii) The information matrix corresponding to parameter α is: $J_0 = E_0[-\frac{\partial^2 \log l(y_t|y_{t-1};\alpha)}{\partial\alpha\partial\alpha'}]$ constrained by the unit mass restrictions on $\pi, \beta_k^*, \gamma_k^*, k = 1,\ldots,K$.

This matrix has a smaller dimension, but is not of full rank due to the lack of identification.

The algorithm introduced above has extended the constrained maximum likelihood approach by adding asymptotically the identification restrictions corresponding to the first-order conditions of the optimization of $\tilde{g}(q,\alpha)$ with respect to q, that are :

$$\frac{\partial \tilde{g}(q,\alpha)}{\partial q} = 0 \Rightarrow q = q(\alpha), \tag{4.12}$$

of a number equal to the degree of underidentification.

These limiting conditions have been replaced by $\frac{\partial \tilde{g}}{\partial q}[q(\hat{\alpha}_T), \hat{\alpha}_T] = 0$ in the IML algorithm and can be expanded in a neighbourhood of $[q(\alpha_0^*) = 0, \alpha_0^*]$. We get :

$$\left[\frac{\partial^2 \tilde{g}}{\partial q \partial q'}[0,\alpha_0^*]\frac{dq}{d\alpha}(\alpha_0^*) + \frac{\partial^2 \tilde{g}}{\partial q \partial \alpha'}(0,\alpha_0^*)\right]\sqrt{T}(\hat{\alpha}_T - \alpha_0^*) \simeq 0,$$

that are the additional asymptotic linear restrictions $D_2'\sqrt{T}(\hat{\alpha}_T - \alpha_0^*) = 0$ on $\sqrt{T}(\hat{\alpha}_T - \alpha_0^*)$ to be taken into account for the computation of the asymptotic variance of $\hat{\alpha}_T$. More precisely,

Proposition 5 : If α_0^* is in the interior of $\mathcal{A}_0$, if the associated π_{0k}^* are all distinct, if assumptions A.1-A.4 and the additional regularity assumptions a.1-a.2 are satisfied, then ;

$$\sqrt{T}(\hat{\alpha}_T - \alpha_0^*) \xrightarrow{d} N(0, J_0^{11}\ J\ J_0^{11}),$$

where J_0^{11} is the North-West block in the block decomposition of the inverse of matrix $\begin{pmatrix} J_0 & D \\ D' & 0 \end{pmatrix}$, $D = (D_1, D_2)$, D_1 defining the $2K+1$ unit mass restrictions on $\pi_k, \beta_k^*, \gamma_k^*, k = 1,\ldots,K$, and D_2 the (asymptotic) linearized restrictions corresponding to the intermediate optimizations in the IML algorithm.

In the expression of the asymptotic variance-covariance matrix, we observe the three main components of the information related with the unconstrained ML of α, the unit mass restrictions D_1 and the restrictions D_2 due to additional intermediate optimization, respectively.

Proof : i) The proof is standard and based on the asymptotic expansion of the first-order conditions on the Lagrangean to account for the equality restrictions (note that the inequality restrictions are not binding if α_0^* belongs in the interior of $\mathcal{A}_0$). These asymptotic expansions are :

$$\begin{pmatrix} J_0 & D \\ D' & 0 \end{pmatrix} \begin{bmatrix} \sqrt{T}(\hat{\alpha}_T - \alpha_0^*) \\ \sqrt{T}(\hat{\lambda}_T - \lambda_0^*) \end{bmatrix} \simeq \begin{bmatrix} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{\partial \log l}{\partial \alpha}(y_t|y_{t-1}; \alpha_0^*) \\ 0 \end{bmatrix}, \tag{4.13}$$

where $\hat{\lambda}_T$ is the associated estimator of the Lagrange multipliers. Because α_0^* is a maximizer of $E_0 \log l(y_t|y_{t-1}, \alpha)$, the normalized score in (4.13) is asymptotically normally distributed with mean zero and variance J.

The result follows whenever the matrix $\begin{pmatrix} J_0 & D \\ D' & 0 \end{pmatrix}$ is invertible. [16]

ii) Let us now discuss this invertibility condition by finding the null space of this matrix, i.e. the solutions θ, λ of the system :

$$\begin{cases} J_0\theta + D\lambda = 0, \\ D'\theta = 0. \end{cases}$$

We know that $D'J_0\theta + D'D\lambda = 0 \Rightarrow \lambda = -(D'D)^{-1}D'J_0\theta$. Then the system in θ only is :

$$\begin{cases} (Id - P)J_0\theta = 0, \\ D'\theta = 0. \end{cases}$$

where P is the orthogonal projector on the space generated by D.

Since the columns of $(Id - P)J_0$ are orthogonal to the columns of D, we see that $\theta = 0$ is the unique solution if and only if : $Rk((Id - P)J_0) = \dim \alpha - \dim q - 1 - 2K$. This statement is Assumption a.2 viii) in online Appendix 2.

QED

Remark 6: The sequence of optimizations cannot be replaced by a penalty term in the objective function. By analogy with the l^1-penalty introduced in LASSO [Tibshirani (1996)], the machine learning literature suggests to use a penalty in the objective function to ensure

[16] We cannot use the usual block formula to compute J_0^{11} [see e.g. Gourieroux, Monfort, Section 10.3.b) because J_0 is not invertible due to the identification issue. However, it is easy to check that a closed form expression of the asymptotic variance of the estimator is :

$$[(Id - P)J_0(Id - P) + P]^{-1}(Id - P)J_0(Id - P)[(Id - P)J_0(Id - P) + P]^{-1},$$

where P is the orthogonal projector on the space generated by D.

the (numerical) convergence of the ML algorithm [see e.g. Kim, Park (2008), Schachtner et al. (2011)], or to circumvent the curse of dimensionality [DePaula et al. (2020), eq. (10). See also Uhlig (2005), Appendix B.2, Mountford, Uhlig (2009) for applications to macro-economics]. In our framework, this penalty function approach (PFA) would lead to an objective function of the form $\log l_T(y;\alpha) + \lambda_T g(\alpha)$, where the tuning parameter λ_T would be an appropriately chosen function of T to ensure the numerical convergence. It is easy to see why this approach would not provide an estimator converging to a benchmark element of the identified set. Indeed, the asymptotic first-order conditions would involve $\frac{\partial \log l_T(y,\alpha)}{\partial \alpha} + \lambda_T \frac{\partial g(\alpha)}{\partial \alpha}$. They are not aligned with the direction allowing to remain in the set, since $\frac{\partial g(\alpha)}{\partial \alpha} = \frac{d\tilde{g}[q(\alpha),\alpha]}{d\alpha}$ differs from $\frac{\partial \tilde{g}(q,\alpha)}{\partial q}$ [see also Ariaz et al. (2018), Section 5, for a critique of PFA in a SVAR model with partial identification].

The asymptotic Gaussian uncertainty is driving all the uncertainties on the true set $\mathcal{A}_0$ of NMF's. More precisely, any other element of $\mathcal{A}_0$ can be written as a deterministic function of α_0: $\xi(q,\alpha_0^*)$, with $q \in Q(\alpha_0^*)$. Then, that element can be estimated by $\xi(q,\hat{\alpha}_T)$, which is a given function of α_T. Therefore, it inherits the asymptotic properties of $\hat{\alpha}_T$: it is consistent of $\xi(q,\alpha_0^*)$, asymptotically normal, and its asymptotic variance-covariance matrix is obtained from the Slutsky formula, whenever q is not on the boundary of $Q(\alpha_0^*)$. If q is on the boundary, its distribution will become truncated normal, and can be easily found by simulation.

Remark 7: The set $\mathcal{A}_0$ is of a large dimension and it is not possible to represent it in a 2 or 3-dimensional space. However it is possible to consider "cuts" of that set obtained by varying a given component of q and setting the other components equal to zero, to examine how the NMF responds to changes in that component. This analysis will depend on the application. Let us consider the static NMF applied to image analysis. The criterion $\det(B'B)$ is likely a measure of contrast and the above approach could be used to find the component q_j that is preferred for changing the contrast of the photo BC' from low to high. Another direction could by used to manage the brightness of the image and so on.

For illustration, let us consider the matrix : $A = \frac{1}{10}\begin{pmatrix} 2 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$. Its elements sum up to one and A defines a joint probability distribution. It is easy to check that $Rk(A) = Rk_+(A) = 2$. This matrix is a mixture of two joint distributions satisfying the independence condition, which can be written in an infinite number of ways. For example, matrix A can be written as :

$$A = \frac{1}{4}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}(3/5, 1/5, 1/5) + \frac{3}{4}\begin{pmatrix} 1/5 \\ 2/5 \\ 2/5 \end{pmatrix}(1/3, 1/3, 1/3),$$

or as :

$$A = \frac{1}{10}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}(1, 0, 0) + \frac{9}{10}\begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}(1/3, 1/3, 1/3).$$

The volumes of the parallelepiped generated by the β^* (resp. γ^*) are 8/25 for the first decomposition and 2/9 for the second decomposition (resp. 8/225 and 2/9). The heterogeneity distribution is less concentrated in the first decomposition than in the second one, its β^* vectors are less colinear and its γ^* vectors are more colinear than in the second one. Note that for $K = 2, \pi_1$ is an homographic function of q_{12} for a given q_{21} (resp. of q_{21} for a given q_{12}) and hence it is monotonic.

Remark 8: The IML approach can be used to derive the lower and upper bounds for partially identified scalar parameters, and to obtain the measures of uncertainty on these bounds [see e.g. Imbens, Manski (2004), Stoye (2009)]. Typical examples are the minimum and maximum values of functions $\sum_{k=1}^{K}(\pi_k \log \pi_k) \det(\tilde{B}'\tilde{B}), \det(\tilde{C}'\tilde{C})$. This procedure is analogous to determining the confidence intervals for average marginal effects in a fixed effects panel logit model [Liu et al. (2021), Davezies et al. (2022)]. Likely, there exists an interval of admissible values of the above uncertainty measures. Such an interval is easy to obtain for the measure of concentration when $K = 2$. However, the lower and upper bounds will be reached on the boundaries of the identified set, for example on the bounds for q_{12}, q_{21} when $K = 2$. Then, the joint asymptotic distribution of these bounds cannot be Gaussian due to the effect of infimum in Proposition 2. Therefore Assumption A.1 i) of normality of the upper and lower bounds in Imbens et al. (2004), Stoye (2009) is not satisfied in our framework. Note that the joint asymptotic distribution of these two bounds is easily derived by simulations and, by construction, we cannot have bound reversal in the estimation.

Remark 9:

The IML approach can also be used to derive a confidence set for $\mathcal{A}_0$ with the correct asymptotic level following Shi and Shum (2015). However this asymptotic confidence set is difficult to represent graphically due to the large dimension of $\mathcal{A}_0$.

4.3.4 Asymptotic distribution of $\hat{A}_T$

The IML approach helps in finding the estimates and confidence intervals of identifiable parameters, especially of the elements a_{ij} of matrix A. In practice, we usually encounter the curse of dimensionality in the unconstrained estimation of A. Moreover, the estimator has to be applied under the constraint of a given non-negative rank : $Rk_+(A) = K$, and the rank-constrained confidence intervals are likely narrower than the unconstrained ones. They can be derived from $\hat{\alpha}_T$ by simulations, given that :

$$a_{ij} \simeq \hat{a} \sum_{k=1}^{K} \hat{\pi}_k \hat{\beta}^*_{ik} \hat{\gamma}^*_{jk} = \hat{a}_{ij}.$$

Asymptotically, $\hat{a}_{ij}$ will converge to the true value $a_{ij,0}$ that does not depend on the choice of α^*_0 benchmark. Similarly, its asymptotic variance-covariance matrix is also independent of this choice, i.e. of the selected function $\tilde{g}$. The additional constraints are only introduced to solve the identification issue.

Proposition 6

The asymptotic distribution of the IML estimator of matrix A does not depend on the benchmark α^*_0 in the identified set, i.e. on the additional optimization criterion.

Proof:

The asymptotic first-order conditions involve $J_0\sqrt{T}(\hat{\theta}_T - \alpha^*_0) + D_1\sqrt{T}(\hat{\lambda}_T - \lambda^*_{10}) + D_2\sqrt{T}(\hat{\lambda}_{2T} - \lambda^*_{20})$ and $D'_1\sqrt{T}(\hat{\alpha}_T - \alpha^*_0) + D'_2\sqrt{T}(\hat{\alpha}_T - \alpha^*_0)$ in the left hand side of system (4.13). Asymptotically, a change of the benchmark modifies the matrix D_2 as well as the associated Lagrange multipliers by linear transformations R, R^{-1}, respectively. Then, the first-order condition provide the same solution for $\sqrt{T}(\hat{\alpha}_T - \alpha^*_0)$ when D_2 is replaced by $\tilde{D}_2 = D_0 R$ and $\hat{\lambda}_{2T} - \lambda^*_{20}$ by $\hat{\tilde{\lambda}}_{2T} - \tilde{\lambda}^*_{20} = R^{-1}(\hat{\lambda}_{2T} - \lambda^*_0)$, where R is invertible. This proves that the asymptotic variance covariance matrix is independent of the choice of the additional optimization criterion.

Q.E.D.

We have introduced an IML estimator of A_0 under the non-negative rank restriction $Rk_+(A_0) = K_0$, derived its asymptotic Gaussian behavior [17] and the expression of the associated efficiency bound. The asymptotic variance-covariance matrix of $\hat{A}_T$ is obtained by applying the Slutsky formula based on the first-order expansions of $a\sum_{k=1}^{K} \pi_k \beta^*_k \gamma^*_k$ in a neighbourhood of $\alpha^*_0 = (a^*_0, \pi^*_{0k}, \beta^*_{0k}, \gamma^*_{0k},\ k = 1, ..., K)$. Then, the asymptotic confidence intervals that are identifiable on identifiable parameter functions of A also have the asymptotic optimality properties.

[17] This asymptotic Gaussian distribution is degenerate because of the reduced rank.

5 Extension to Nonparametric Identification

5.1 The Identified Set

The proof of Proposition 1 is valid in a nonparametric framework. Let us consider a pair of real variables (X, Y) with a continuous joint distribution on a product of intervals and assume a positive pdf $f(x; y)$ on that interval. We can be interested in a decomposition :

$$f(x, y) = \sum_{k=1}^{K} \pi_k \beta_k(x) \gamma_k(y), \tag{5.1}$$

where (π_k) defines the latent heterogeneity distribution, $\beta_k(x), \gamma_k(y)$ some pdf's for x and y, respectively. This is a mixture model in which each joint distribution in the mixture satisfies the independence condition [see Compiani, Kitamura (2016) for mixtures in Econometrics] [18].

The proof remains valid if functions $\beta_1, \ldots, \beta_K$ (resp. $\gamma_1, \ldots, \gamma_K$) are a.s. linearly independent. Since the model is already normalized, Proposition 1 becomes :

Proposition 1' : Let us consider a true factorization of the joint pdf :

$$f_0(x, y) \;\; = \;\; \sum_{k=1}^{K} b_{0k}(x) \gamma_{0k}(y) = B_0(x) C_0(y),$$

where the terms b_{0k} are positive densities (not necessarily with unit mass) and γ_{0k} are probability densities. Then, the other observationally equivalent decompositions are $B(x) = B_0(x)Q, C(y) = C_0(y)(Q')^{-1}$, where the matrix Q is invertible, with unitary diagonal elements and such that :

$$B_0(x)Q \geq 0, \text{ a.s.}, \;\; C_0(y)(Q')^{-1} \geq 0 \quad \text{a.s.}$$

The decomposition (5.1) of the joint pdf has alternative interpretations. For instance, the conditional p.d.f. can be written as:

$$f(y|x) = \frac{f(x, y)}{f(x)} = \sum_{k=1}^{K} \frac{b_{0k}(x)}{f(x)} \gamma_{0k}(y) \equiv \sum_{k=1}^{K} b^*_{0k}(x) \gamma_{0k}(y).$$

[18] This condition can be written for more than two variables, so that the mixture becomes identifiable and easier to analyse [Hall, Zhou (2003), Kasahara, Shimotsu (2009)]. A suitable notion of nonnegative rank has not yet been introduced for 3- or 4-entry tables.

Therefore, it is equivalent to impose a reduced rank condition on the joint distribution, or on the conditional distribution. This problem is considered in Henry , Kitamura, Salanie (2014), where the pair of variables is denoted by Y, W, the representation is written conditionally on a third variable X, and the independence condition on the elements of the mixtures is called the exclusion restriction [see also Compiani, Kitamura (2016)].

5.2 Nonparametric Identifying Maximum Likelihood

The estimation approach outlined in Section 4 can be extended to the functional parameter framework, with functional parameters $b_k, k = 1, \ldots, K$, scalar parameters $\pi_k, k = 1, \ldots, K$, and additional parameters in vec^*Q.

Let us assume i.i.d. observations on $(X_i, Y_i), i = 1, \ldots, n$. Then, a kernel based AML can be applied with the objective function :

$$\sum_{i=1}^{n} \{K(\frac{x_i - x}{h_n}) K(\frac{y_i - y}{h_n}) \log \sum_{k=1}^{K} [b_k(x)\gamma_k(y)]\},$$

where $K(.)$ denotes a kernel, h_n the bandwidth and $\int \gamma_k(y)dy = 1, \ k = 1, ..., K$. The maximisation with respect to $b_k(x), \gamma_k(y), k = 1, \ldots, K$ provides the functional estimators of $b_k(.), \gamma_k(.), \ k = 1, \ldots, K$, and then functional estimators of $\pi_k, \beta_k(.), \gamma_k(.)$.

Next, due to the identification issue, an additional optimisation has to be performed to fix a mixture in the identified set. Criteria equivalent to the criteria introduced in Section 4.3.2 can be used, either the concentration criterion $\sum_{k=1}^{K} (\pi_k \log \pi_k)$, or the alternative criterion $\det \Gamma$, with

$$\Gamma = \int \gamma(y)\gamma'(y)dy.$$

The analysis of the asymptotic properties of this functional estimator approach are out of the scope of this paper and left for future research [19].

6 Concluding Remarks

Although the nonnegative matrix factorization (NMF) is a well-known technique of dimension reduction for nonnegative matrices, it is used in the absence of an associated probability model

[19] In this framework the benchmark is a functional parameter, a case that is not included in Shi, Shum (2015).

for the observed data. Our paper fills this gap by considering a structural dynamic network model. We suggest new estimation methods for the set of NMF's and derive the asymptotic distribution of the estimated set and of specific elements of that set. Moreover, we provide a ML estimator of the non-negative matrix A_0 under a given non-negative rank constraint. The proposed approach is related to the nonparametric identification in mixture models.

Our approach can be used in a variety of applications with a lack of local identifiability. When an element of the identified set is characterized as a solution of an auxiliary optimization, the identified set can be parametrized given this element considered as a new origin. In practice, the dimension of the parametric identified set can be very large and cannot be represented in a low-dimensional figure. However, it is possible to illustrate various elements or cuts of that set, which have structural interpretations and are easier to represent and discuss.

A typical partial identification of the same type is encountered in the analysis of the effects of monetary policy shocks under sign restrictions on the impulse response functions [Uhlig (2005), Baumeister, Hamilton (2015), Ariaz, Rubio-Ramirez, Waggoner (2018), Granzeria et al. (2018)]. For example, in a VAR(1) model $Y_t = \Phi Y_{t-1} + u_t$, with $u_t \sim N(0, \Sigma)$ and $\Sigma = AA'$. The matrix A is not identifiable. The identified set can be reduced by imposing a restriction so that a "monetary policy impulse vector (a column of A) implies negative responses on prices and nonborrowed reserves and positive responses on federal funds rate, at all horizons" [Assumption A.1, Uhlig (2005)]. Then, there is a reduced identified set of these impulse vectors that can be easily parametrized. The idea of determining first the specific elements of the identified set by means of an additional optimization, before performing an extreme bounds analysis in the spirit of Leamer (1983) appears in Uhlig (2005), page 388 and Appendix B.2.

The standard factor analysis of time series by SVD is commonly followed by an interpretation of the dynamic factors. The factorial directions are projected on some observable time series to provide economic or financial interpretations of the dynamic factors (the so-called mimicking factors). A similar approach can be applied in our framework with partial identification. This may lead to selecting the most interpretable β_k's or $\gamma_k's$.

References

[1] Anderson, T., (1963) : "Asymptotic Theory for Principal Component Analysis", Ann. Math. Statist., 34, 122-148.

[2] Anderson, T., and H., Rubin (1956) : "Statistical Inference in Factor Analysis", in Proc. 3rd Berkeley Symp. Mathematical Statistics and Probability, ed. J. Neyman, Vol 5, 111-150, Berkeley, Univ. of California Press.

[3] Andrews, D and G. Soarez (2010): "Inference for Parameters Defined by Moment Inequalitis Using Generalized Moment Selection", Econometrica, 78, 119-157.

[4] Ariaz, J., Rubio-Ramirez, J. and D. Waggoner (2018): "Inference Based on Structural Vector Autoregressions Identified with Sign and Zero Restrictions: Theory and Applications", Econometrica, 86, 685-720.

[5] Barth, N. (1999) : "The Gramian and K-volume in N-Space : Some Classical Results in Linear Algebra", Journal of Young Investors, 2.

[6] Baumeister, C. and J. Hamilton (2015): "Sign Restrictions, Structural Vector Autoregressions and Useful Prior Information", Econometrica, 83, 1963-1999.

[7] Berman, A., and R., Plemmons (1994) : "Nonnegative Matrices in the Mathematical Sciences", Philadelphia, SIAM, Elsevier.

[8] Blume, L., Brock, W., Durlauf, S., and Y., Ionnides (2011) : "Identification of Social Interactions", Handbook of Social Economics, Vol 1, 853-964.

[9] Blume, L., Brock, W., Durlauf, S., and R., Jayaraman (2015) : "Linear Social Interaction Models", Journal of Political Economy, 123, 444-496.

[10] Bramoulle, Y., Djebbari, H., and B., Fortin (2009) : "Identification of Peer Effects Through Social Networks", Journal of Econometrics, 150, 41-55.

[11] Brie, D. (2015) : "On the Uniqueness and Admissible Solutions of Nonnegative Matrix Factorization", Winter School, Search for Latent Variables : ICA's, Tensors and NMF, Villard de Lans.

[12] Brock, W., and S., Durlauf (2010) : "Adoption Curves and Social Interactions", Journal of the European Economics Association, 8, 235-251.

[13] Cai, J., Yang, D. Zhu, W. Shen, H. and L. Zhao (2022): "Network Regressions and Supervised Centrality Estimation", DP. University of Pennsylvania.

[14] Canay, I., and A. Shaikh (2017): "Practical and Theoretical Advances in Inference for Partially Identified Models", in Advances in Economics and Econometrics, B. Homore, A. Pakes, M. Piazzesi, and L. Samuelson eds., Vol 2. Chap 9, 271-306, Cambridge University Press.

[15] Chen, X. , Christensen, M. and E. Tamer (2018): "Monte-Carlo Confidence Sets for Identified Sets", Econometrica, 86, 1965-2018.

[16] Chen, M., Fernandez-Val, I., and M., Weidner (2021) : "Nonlinear Factor Model for Network and Panel Data", Journal of Econometrics, 220, 296-324.

[17] Chernozhukov, V., Hong, H., and E., Tamer (2007) : "Estimation and Confidence Regions for Parameter Sets in Econometric Models", Econometrica, 75, 1243-1284.

[18] Cohen-Cole, E., Kirilenko, A. and E., Patacchini (2014) : "Trading Networks and Liquidity Provision", Journal of Financial Economics, 113, 235-251.

[19] Collins, M., Dasgupta, S., and R., Shapiro (2002) : "A Generalization of Principal Component Analysis to the Exponential Family", in Advances in Neural Information Processing Systems, 14, Dietterich, T., Becker, S., and Z., Ghahramani eds., MIT Press, 617-624.

[20] Compiani, G., and Y., Kitamura (2016) : "Using Mixtures in Econometric Models : A Brief Review and Some New Results", Econometrics Journal, 19, 95-127.

[21] Davezies, L., D'Haultfoeuille, and L., Laage (2022) : "Identification and Estimation of Average Marginal Effects in Fixed Effects Logit Models", arXiv 2105.00879.

[22] De Giorgi, G. Pellizzari, M., and S. Redaelli (2010) : "Identification of Social Interactions Through Partially Overlapping Peer Groups", American Economic Journal : Applied Economics, 2, 241-275.

[23] De Paula, A. (2017) : "Econometrics of Network Models", in Advances in Economics and Econometrics, Eds, Honore, B., Pakes, A., Piazzesi, M, and L., Samuelson, Cambridge Univ., Press Chapter 8.

[24] De Paula, A., Rasul, I., and P., Souza (2020) : "Identifying Network Ties from Panel Data : Theory and an Application to Tax Competition", DP University College, London.

[25] Donnet, S. and S. Robin (2021): "Accelerating Bayesian Estimation for Network Poisson Models Using Frequentist Variational Estimates", JRSS Series C, 70, 858-885.

[26] Fahrenwaldt, M., Weber, S., and K., Weske (2018) : "Pricing of Cyber Insurance Contracts in a Network Model", ASTIN Bulletin, 48, 1175-1218.

[27] Fernandez-Val, I., and M., Weidner (2016) : "Individual and Time Effects in Nonlinear Panel Models with Large N, T", Journal of Econometrics 192, 291-312.

[28] Gill, P., Murray, W., and M., Saunders (2002) : "SNOPT : An SQP Algorithm for Large Scale Constrained Optimization", SIAM J. Optim., K, 979-1006.

[29] Gourieroux, C., and J., Jasiak (2022) : "SIR Model with Multiple Transmission", CREST DP.

[30] Gourieroux, C., and A., Monfort (1995) : "Statistics and Econometric Models", Vol 1, Cambridge University Press.

[31] Gourieroux, C., Monfort, A., and E., Renault (1990) : "Biilinear Constraints : Estimation and Test", Essays in Honor of Edmond Malinvaud, Empirical Economics, MIT Press, 166-191.

[32] Granzeria, E., Moon, H. and F. Schorfheide (2018): 'Inference for VARs Identified with Sign Restrictions", Quantitative Economics, 9, 1987-2111.

[33] Hall, P., and X., Zhou (2003) : "Nonparametric Estimation of Component Distributions in a Multivariate Mixture", Annals of Statistics, 31, 201-224.

[34] Henry, M., Kitamura, Y., and B., Salanie (2014) : "Partial Identification of Finite Mixtures in Econometric Models", Quantitative Economics, 5, 123-144.

[35] Hillairet, C., and O., Lopez (2020) : "Propagation of Cyber Incidents in an Insurance Portfolio : Counting Processes Combined with Comportmental Epidemiological Models", Hal-02564462.

[36] Hong, H., and J., Xu (2014) : "Count Models of Social Networks in Finance", DP Princeton Univ.

[37] Imbens, G., and C., Manski (2004) : "Confidence Intervals for Partially Identified Parameters", Econometrica, 72, 1845-1857.

[38] Kasahara, H., and K., Shimotsu (2009) : "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices", Econometrica, 77, 135-175.

[39] Kim, H., and H., Park (2007) : "Sparse Nonnegative Matrix Factorization via Alternating Non-Negativity Constrained Least Squares for Microarray Data Analysis", Bioinformatics, 23, 1495-1505.

[40] Kim, H., and H., Park (2008) : "Sparse Nonnegative Matrix Factorization for Clustering", Georgia Tech., GT-CSE-08-01.

[41] Kline, B. and E. Tamer (2022): "Recent Developments in Partial Identification", D.P. Harvard University.

[42] Laurberg, H., Christensen, M., Plumbey, M., Hansen, L., and S., Jensen (2008) : "Theorems on Positive Data : On the Uniqueness of NMF", Computational Intelligence and Neuroscience, ID 764206.

[43] Leamer, E. (1983): "Let's Take the Con out of Econometrics", American Economic Review, 73, 31-43.

[44] Lee, L., Liu, X., and S., Lin (2010) : "Specification and Estimation of Social Interaction Models with Network Structure, Contextual Factors, Correlation and Fixed Effects", Econometrics Journal, 13, 145-176.

[45] Lee, D., and H., Seung (1999) : "Learning the Parts of Objects by Nonnegative Matrix Factorization", Nature, 401, 788-791.

[46] Lewbel, A. (2019) : "The Identification Zoo : Meanings of Identification in Econometrics", Journal of Economic Literature, 57, 835-903.

[47] Liu, X. (2005) : "Global Convergence on an Active Set SQP for Inequality Constrained Optimization", Journal of Computational and Applied Mathematics, 180, 201-211.

[48] Liu, L., Poirier, A., and J., Shin (2011) :" Identification and Estimation of Average Partial Effects in Semi-Parametric Binary Response Panel Models", arXiv : 2105.12891.

[49] Manski, C. (1993) : "Identification of Endogenous Social Effects : The Reflection Problem", The Review of Economic Studies, 60, 531-542.

[50] McCullagh, P., and J., Nelder (1989) : "Generalized Linear Models", 2nd edition, London, Chapman and Hall.

[51] Meyer, C. (2000) : "Matrix Analysis and Applied Linear Algebra", SIAM, Philadelphia.

[52] Mountford, A., and H. Uhlig (2009): "What Are the Effects of Fiscal Policy Shocks", Journal of Applied Econometrics, 24, 960-992.

[53] Nocedal, J., and S., Wright (2006) : "Numerical Optimization", 2^{nd} edition, Springer, Berlin.

[54] Schachtner, R., Pappel, G., and E., Lang (2011) : "Toward Unique Solutions of Nonnegative Matrix Factorization Problems by a Determinant Criterion", Digital Signal Processing, 21, 528-534.

[55] Shi, X. and M. Shum (2015): "Simple Two-Stage Inference for a Class of Partially Identified Models", Econometric Theory, 31, 493-520.

[56] Stoye, J. (2009) : "More on Confidence Intervals for Partially Identified Parameters", Econometrica, 77, 1299-1315.

[57] Tibshirani, R. (1996) : "Regression Shrinkage and Selection via the, LASSO", JRSS, B, 58, 267-288. -

[58] Tipping, H., and C., Bishop (1999) : "Probabilistic Principal Component Analysis", JRSS, B, 61, 611-622.

[59] Uhlig, H. (2005): "What are the Effects of Monetary Policy on Output? Results from an Agnostic Identification Procedure", Journal of Monetary Economics, 52, 381-419.

Online Appendix 1

Comparison of the Decompositions

1. The NMF

After the transformation the new NMF is :

$$\begin{aligned} A &= \tilde{\beta}_1\tilde{\gamma}_1' + \tilde{\beta}_2\tilde{\gamma}_2', \\ \text{with } \tilde{\beta}_1 &= \beta_1 + q_{21}\beta_2 = q_{12}\beta_1 + \beta_2, \\ \tilde{\gamma}_1 &= \frac{1}{1-q_{12}q_{21}}(\gamma_1 - q_{12}\gamma_2), \tilde{\gamma}_2 = \frac{1}{1-q_{12}q_{21}}(-q_{21}\gamma_1 + \gamma_2). \end{aligned}$$

It is easily checked that $A = \beta_1\gamma_1' + \beta_2\gamma_2'$.

2. The decomposition (2.13).

To get the new decomposition (2.13) :

$$A = a(\tilde{\pi}\tilde{\beta}_1^*\tilde{\gamma}_1^{*'} + (1-\tilde{\pi})\tilde{\beta}_2^*\tilde{\gamma}_2^{*'}],$$

the new factorial directions have to be normalized with components summing up to one. We get :

$$\begin{aligned} \tilde{\beta}_1^* &= (\beta_1 + q_{21}\beta_2)/(\beta_1'e + q_{21}\beta_2'e), \\ \tilde{\beta}_2^* &= (q_{12}\beta_1 + \beta_2)/(q_{12}\beta_1'e + \beta_2'e), \\ \tilde{\gamma}_1^* &= (\gamma_1 - q_{12}\gamma_2)/(\gamma_1'e - q_{12}\gamma_2'e), \\ \tilde{\gamma}_2^* &= (-q_{21}\gamma_1 + \gamma_2)/(-q_{21}\gamma_1'e + \gamma_2'e), \\ \tilde{\pi}/(1-\tilde{\pi}) &= (\beta_1'e + q_{21}\beta_2'e)(\gamma_1'e - q_{12}\gamma_2'e)/[(q_{12}\beta_1'e + \beta_2'e)[-q_{21}\gamma_1'e + \gamma_2'e)]. \end{aligned}$$

Online Appendix 2

Additional Assumptions for Asymptotic Results

1. Consistency

We provide below a set of additional assumptions a.1 to get the consistency. They require some uniform convergence of the objective function on the set $\mathcal{A}^*$ of all possible α on which the optimization is performed.

Assumption a.1 :

i) The set $\mathcal{A}^*$ is compact.
ii) $\log l(y_t|y_{t-1};\alpha)$ is integrable for all $\alpha \in \mathcal{A}^*$.
iii) $\mathcal{A}_0 \subset \mathcal{A}^*$.
iv) Uniform convergence of the objective function :

$$\sup_{\alpha\in\mathcal{A}^*} |\frac{1}{T}\sum_{t=1}^{T} \log l(y_t|y_{t-1};\alpha) - E_0 \log l(y_t|y_{t-1};\alpha)|$$

$$= 0_P(1/\sqrt{T}),$$

where E_0 is the expectation with respect to the stationary distribution of (Y_{t-1}, Y_t).
v) $\lim_{p\to\infty} \frac{1}{T}\sum_{t=1}^{T} \log l(y_t|y_{t-1};\hat{\alpha}_T(\alpha_0,p)) = \max_{\alpha\in\mathcal{A}^*} \frac{1}{T}\sum_{t=1}^{T} \log l(y_t|y_{t-1};\alpha)$.

The three first conditions are standard and used to prove the convergence of the set of solutions of the finite sample optimisations to $\mathcal{A}_0$. They imply condition C_1 on Chernozhukov et al. (2007) for instance (See also this reference for a proof, in which, in our framework, the objective function is the log-likelihood function instead of a moment criterion function). This is the convergence result in Proposition 3. Proposition 4 follows since this convergence of sets is uniform.

The last condition iv) is usually not introduced. It concerns the algorithm used to approximate solutions of the finite sample optimisation problems. This condition explains why we have introduced stronger conditions as in Assumption A.4 on the concavity of the log-likelihood function.

By construction the domain for $\pi, \beta_k^*, \gamma_k^*, k = 1, \ldots, K$ is compact and its bounds as $\pi_k = 0$, for some k, or $\beta_k = 0$, for some k cannot be reached due to the rank condition, i.e. Assumptions A.2 and A.3. Therefore assumption a.1 i) concerns mainly scalar parameter a.

2. Asymptotic Normality

When T tends to infinity, the estimator $(\hat{\alpha}_T, \hat{q}_T = q(\hat{\alpha}_T))$ will tend to $(\alpha_0^*, 0)$. Let us assume :

Assumption a.2 :

i) The true set $\mathcal{A}_0$ has a non-empty interior and α_0^* is in the interior of the true set $\mathcal{A}_0$.

ii) 0 is in the interior of the set $Q(\alpha_0^*)$ of admissible values of q constructed from α_0^*.

iii) The log-likelihood function is twice continuously differentiable with respect to α.

iv) The additional objective function $\tilde{g}$ is continuously differentiable with respect to q and continuously cross differentiable with respect to q and α.

v) The function $q(.)$ exists and is continuously differentiable

vi) The score $\dfrac{\partial \log l}{\partial \alpha}(y_t|y_{t-1};\alpha)$ has second-order moments.

vii) The Hessian $\dfrac{\partial^2 \log l}{\partial \alpha \partial \alpha'}(y_t|y_{t-1};\alpha)$ has first-order moments.

viii) The matrix $(Id - P)J_0$ has rank $\dim \alpha - \dim q - 1 - 2K$.

Let us discuss these additional assumptions. Condition $a.2$ i) eliminates the case $K = 1$, when the NMF is point identified and the standard asymptotic theory applies. This is assumption (4^*) in Shi, Shum (2015), Theorem 2.1. Then, the rank condition in their assumption (4) is automatically satisfied in our framework by assumptions A-2-A.3 and a.2 viii). Their condition (***) in Theorem 3.1 is automatically satisfied in our framework of maximum likelihood estimation.