

Inference on Nonnegative Matrix Factorization from Identifying Maximum Likelihood

Gourieroux*, C. and J. Jasiak†

December 30, 2025

Abstract

This paper examines nonlinear dynamic models where the main parameter of interest is a nonnegative matrix characterizing the network (contagion) effects that admits a reduced-rank Nonnegative Matrix Factorization (NMF). We develop a new probabilistic NMF inference approach based on a novel Identifying Maximum Likelihood (IML) method for consistent estimation of the identified set of admissible NMF's and derive the asymptotic distribution of this random set. We also propose a maximum likelihood estimator of the network matrix for a given nonnegative rank, and derive its asymptotic distribution and the associated efficiency bound.

Keywords: Network, Nonnegative Matrix Factorization (NMF), Set Identification, Random Set, 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 C. Bontemps, X. D'Haultfoeuille, R. Engle, M. Henry, A. Monfort and the participants of the CREST Financial Econometrics seminar, CIREQ Econometrics Conference in Montreal 2024, African Meeting of the Econometric Society, TSE Econometrics Seminar and Panel Data Conference in Orleans for helpful comments.

1 Introduction

A network is commonly characterized by a non-negative matrix A of dimension (n, m) . There exists a large literature on network models and different assumptions are imposed on the non-negative matrix A . These assumptions concern either the adjacency matrices with 0 - 1 elements, or interaction matrices which are supposed to have positive elements and "reduced rank". In applications to contagion, transmission, social interaction, and spillover effects, the elements of each row of matrix A can be additionally constrained to sum up to one. Then, each row of matrix A is interpreted as a conditional probability distribution and matrix A is called a transition, or migration matrix.

Depending on the objective of analysis, network models are either prediction-oriented, or structural. The prediction-oriented models are used for image and facial recognition, text analysis, environmetrics, and genomics. In these applications, the dimensions n and m of matrix A are very large, and the primary objective of the analysis is to reduce this high dimensionality. Then, adjacency matrices with small numbers of links, i.e. elements equal to 1, can be used to obtain sparse matrices A . Alternatively, interaction matrices can be chosen so that the non-negative matrix factorization (NMF) of matrix A is of a small reduced rank [Berman and Plemmons (1994)].

The aim of this paper is to fill the gap between the prediction-oriented and structural analysis by introducing the Identifying Maximum Likelihood (IML) estimation method, that provides the Maximum Likelihood (ML) estimator of the identified set of admissible NMFs, as well as ML estimator of matrix A , under a rank constraint for an extended class of dynamic probabilistic network models.

We introduce a class of dynamic parametric models with non-negative interaction matrices of a reduced rank. It includes the static models used in machine learning, dynamic panel models for individual qualitative variables, and multivariate dynamic Poisson models with contagion used in epidemiology as special cases [Gourieroux and Lu (2019), (2023), Lu et al. (2024)]. In the latter one, the NMF of matrix A may involve not only the principal directions, but also a latent heterogeneity distribution. In general, there exist multiple admissible factorizations of a given true network matrix A_0 . We derive the formulas of the identified set of admissible NMFs for non-

negative ranks $K = 1$ and $K = 2$ of matrix A , and show how the identified set can be parametrized in the general case $K \geq 2$. Our main contribution is the Identifying Maximum Likelihood (IML) approach which is introduced to estimate the identified 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 and/or the least collinear principal directions is considered. Moreover, we propose a maximum likelihood estimator of the identified non-negative matrix A , given its non-negative rank, and derive the efficiency bound for this rank-constrained parameter matrix estimator.

The paper is organized as follows. In Section 2, we describe the class of dynamic parametric models with interaction matrices of reduced rank. Section 3 discusses the identification of the NMF of matrix A . Statistical inference is developed in Section 4. An illustration is provided in Section 5. Section 6 concludes. The list of additional regularity assumptions for asymptotic statistical inference and additional estimation results are given in on-line appendices.

2 Parametric Model with Interaction Matrix

2.1 The model

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

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

where A is an unknown non-negative matrix $A \geq 0$ of dimension (n, m) , which has non-negative entries. We know that a non-negative matrix A can be factorized and written as :

$$A = BC', \quad (2.2)$$

where B (resp. C) is of dimension (n, K) [resp. (m, K)] and B and C are nonnegative : $B \geq 0, C \geq 0$. Among the multiple non-negative factorizations available, some correspond to the minimal order K , called the non-negative rank of matrix A and

denoted by $Rk_+(A)$. The non-negative rank of A is always larger or equal to the rank of A .

Specifically, a non-negative matrix factorization (NMF) can be written as follows. Let $\beta_k, k = 1, \dots, K$ (resp. $\gamma_k, k = 1, \dots, K$) denote the columns of B (resp. C). These columns define the factorial directions. We have :

$$B = (\beta_1, \dots, \beta_K), C = (\gamma_1, \dots, \gamma_K), \quad (2.3)$$

and then :

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

This is a decomposition of A into a sum of K non-negative matrices of rank 1.

In structural models, we may be interested in finding the true value A_0 , as well as a true NMF : $B_0 C_0' = \sum_{k=1}^{K_0} \beta_{0,k} \gamma_{0,k}'$, that generates A_0 .

We make the following assumptions:

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.

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 maximization $\max_A E_0 \log l(Y_t | Y_{t-1}; A)$ with respect to the set of non-negative matrices, has a unique solution $A = A_0$; iii) The vectors $\beta_{0,k}, \gamma_{0,k}, k = 1, \dots, K$, have strictly positive entries.

Remark 1 : This analysis can be easily extended to a) parametric conditional models including observed exogenous variables X_t , or b) additional parameters θ in models of the type: $l(y_t | \underline{y}_{t-1}, \underline{x}_t) = l(y_t | y_{t-1}, x_t; A_0, \theta_0)$, and c) dynamic panel models with covariates.

2.2 Applications

Our objective is to extend the NMF to applications to which the principal component analysis (PCA) [Tipping and Bishop (1999)] is not applicable because it does not ensure the positivity condition. For example, the models for facial recognition, epidemiology, volatility, credit risk and cyber risk analysis have to account for the non-negativity of observations Y_t in practice.

2.2.1 Static Model

The static models assume that observations $Y_t, t = 1, \dots, T$, are independent and identically distributed (i.i.d.). These models are commonly used for image analysis [see Paatero and Tapper (1994) for the introduction of NMF]. NMF started to be extensively studied following Lee and Seung (1999) and their application to learning the parts of objects under a non-probabilistic approach. In image analysis, the observations are matrices $Y_t = (Y_{i,j,t})$, where (i, j) denote the coordinates of a point in picture t and $Y_{i,j,t}$ is the pixel intensity associated with coordinates $i = 1, \dots, n, j = 1, \dots, m$ and picture t . The pixel intensity can be measured on either a discrete, or continuous scale. Then, the model:

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

is based on a family of probability density functions (p.d.f.) $f(y; \lambda)$ with a non-negative y and a non-negative scalar parameter λ . For the Poisson and exponential p.d.f., the model in (2.5) simplifies to a generalized linear model (GLIM) [McCullagh and Nelder (1989)].

Static models are also applied to data on a set of individuals $i = 1, \dots, L$ where at time t we observe the number $y_{i,j,t}$ of messages sent by individual i to individual j . Alternatively, the i.i.d. observations y_t can be matrices containing the investments of bank i in industrial sector j at time t , gravity matrices summarizing international trades of countries [Chen et al. (2021), Section 6], or matrices of numbers of stocks of firms head-quartered in city $j, j = 1, \dots, m$, held by mutual fund manager $i, i = 1, \dots, n$ at time t [Hong and Xu (2014)].

2.2.2 Dynamic Qualitative Panel

Let us consider a panel of L individuals, $l = 1, \dots, L$. Each individual is characterized by a qualitative state $i = 1, \dots, n$, that can be observed at any time t . The qualitative individual histories can be quantified and represented by n -dimensional vectors $(Y_{l,t}, t = 1, \dots, T)$, $l = 1, \dots, L$, where $Y_{l,t}$ has $\{0, 1\}$ entries that sum up to 1. The dynamic model can be defined by assuming that :

i) the individual histories are independent from one another; 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 :

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

Then, the sequence of multivariate counts $(Y_t, t = 1, \dots, T)$ is a Markov process with the conditional p.d.f. obtained from the convolute of different multinomial distributions :

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

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, individuals are replaced by homogeneous segments that diminishes the dimensionality of the network. For example, the corporates can be replaced by industrial sectors, households by age categories, or animals by species [Donnet and 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 the same dimension n , and both vectors y and θ are non-negative. Then, a dynamic model for (y_t) can be defined as :

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

where the contagion matrix A is non-negative and of dimension (n, n) . This model can be easily extended to include an intercept μ , i.e. $l(y_t|y_{t-1}; A) = l(y_t, Ay_{t-1} + \mu)$.

i) When this parametric model represents the dynamics of n independent Poisson variables, we obtain a multivariate Autoregressive Conditional Poisson (ACP) model

with :

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

where a_i is the i^{th} row of matrix A and e is a vector with unitary elements [Cameron and Trivedi (2008), Section 7.5, eq. 7.42, Fokianos (2024)] .

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

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

This dynamic model can be used to study joint evolutions of the gross domestic product in a set of n countries. Then, matrix A is unobserved and needs to be estimated under some mild constraints to provide a proxy of the international trading network. It can also be used for joint analysis of observed (implied or realized) daily volatilities, with the individual index i assigned either to the stocks, or hours of a trading day, in the spirit of multiplicative volatility models [Engle and Rangel (2008), Hafner and Linton (2010)].

Remark 2 : The dynamic model (2.8) can be extended to $l(y_t|y_{t-1}; A) = l(y_t; A z_{t-1})$, where z_{t-1} is a nonnegative vector function of y_{t-1} . Such a transformation appears in structural models used in epidemiology and other applications such as the analysis of cyberattacks [Fahrenwaldt et al. (2018), Lu et al. (2024)], adoption of new technologies [Brock and Durlauf (2010)] 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. Gouriéroux and Lu (2023)].

2.3 Latent Heterogeneity and Ranking

Taking into account the non-negativity condition allows us to interpret and visualize the networks. Let us now introduce a normalized NMF.

2.3.1 Alternative parametrization

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

$$A = a \sum_{k=1}^K \pi_k \beta_k^* \gamma_k^{*'}, \quad (2.11)$$

where $a = e' A e = \sum_i \sum_j a_{i,j} > 0$, $\pi_k \geq 0, k = 1, \dots, K$, with $\sum_{k=1}^K \pi_k = 1$, $\beta_k^* \geq 0, \gamma_k^* \geq 0, k = 1, \dots, K$, with $\beta_k^{*'} e = \gamma_k^{*'} e = 1, k = 1, \dots, K$.

In this decomposition $\pi = (\pi_1, \dots, \pi_K)'$, $\beta_k^*, \gamma_k^*, k = 1, \dots, K$ 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 related to the literature on partial identification of finite mixtures [see Hall and Zhou (2003), Kasahara and Shimotsu (2009), Henry et al. (2014), and Online Appendix a.5].

2.3.2 Rankings

The alternative parametrization (2.11) leads to a dynamic model of count variables in epidemiology, such as the dynamic contagion model where the components of y_t are counts of infected individuals in L homogenous segments of the population. In addition, let us assume that:

$$E_{t-1} y_t = A y_{t-1}. \quad (2.12)$$

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

$$E_{t-1} y_{i,t} = \sum_{j=1}^L a_{ij} y_{j,t-1} = a \beta_{1i}^* \sum_{j=1}^L \gamma_{1j}^* y_{j,t-1}. \quad (2.13)$$

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 . Therefore the segments $i = 1, \dots, L$ can be ranked with respect to their vulnerability β_{1i}^* and their infectiosity γ_{1i}^* .

When K is larger than or equal to 2, the segments are heterogeneous with latent heterogeneity. Its distribution is denoted by π . Then, the segments can be ranked

with respect to their vulnerability ratings β_{ki}^* , $k = 1, \dots, K$, and infectiosity ratings $\gamma_{k,i}^*$, $k = 1, \dots, K$. The potential interpretations of parameters β^* , γ^* , π depend on the application of interest. In the analysis of internet diffusion of messages, parameters β^* (resp. γ^*) can be used for ranking the 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, in cyber risk networks for ranking of firms by vulnerability and hackers by their impact.

3 Identification of the Nonnegative Factorization

In general, the true NMF is not point-identified. This section discusses the identification issues and derives the parametric representation of the identified set.

3.1 The general framework

The parametric model depends on the non-negative matrix factorization :

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

where K denotes the non-negative rank and β_k, γ_k' s are the factorial directions. In practice, the structural parameters are $K, \beta_k, \gamma_k, k = 1, \dots, K$. 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 , or after rescaling by a positive scalar, i.e. by replacing β_k, γ_k by $\sigma_k \beta_k$, and γ_k / σ_k , respectively, for a positive scalar σ_k . This identification issue is easily solved by a normalization. The more complicated identification issues arise when $K \geq 2$ and are discussed below under the following assumption :

Assumption A.3 : The non-negative 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 the following equivalent forms :

Assumption A.3 is satisfied \Leftrightarrow the vectors β_1, \dots, β_K are linearly independent and the vectors $\gamma_1, \dots, \gamma_K$ are linearly independent $\Leftrightarrow B'B$ and $C'C$ are invertible³.

Assumption A.3 implies that the matrices $\beta_k \gamma_l', k, l = 1, \dots, K$ 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 non-negative rank and the rank of A_0 are known. Let us consider another factorization of matrix A without taking into account the non-negativity conditions on β_k, γ_k' s. Since the range of A (resp. A') is the space spanned by β_1, \dots, β_K (resp. by $\gamma_1, \dots, \gamma_K$), an alternative factorization is : $A = B G H' C' = \tilde{B} \tilde{C}'$, where $\tilde{B} = B G, \tilde{C}' = C' H$ and G, H are invertible matrices⁴. Moreover, we have : $B G H' C' = B C'$, and by Lemma 1 we deduce that $H' = G^{-1}$. Therefore we get : $\tilde{B} = B G, \tilde{C}' = C'(G')^{-1}$.

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) invertible matrix with diagonal elements equal to 1. Then, $(G')^{-1} = (Q')^{-1} \text{diag } (1/\sigma)$. By taking into account the non-negativity conditions, we get a constructive characterization of the identified set :

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

$$\tilde{B} = B_0 Q \text{diag } \sigma, \tilde{C}' = C_0 (Q')^{-1} \text{diag } (1/\sigma), \sigma_k > 0, \forall k = 1, \dots, K,$$

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

$$B_0 Q \geq 0, C_0 (Q')^{-1} \geq 0.$$

³The NMF representation is different from the Singular Value Decomposition (SVD) of matrix A . In SVD the identification issue is 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 non-negative. The orthogonality condition would imply $\beta_{ik} \beta_{il} = 0, \forall i$, and contradict assumption A.1.

⁴The columns of B and the columns of \tilde{B} are two bases of the range of AA' equal to the range of A . Therefore they are in a one-to-one relationship represented by an invertible matrix G .

The non-negative 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.

Proposition 1 shows that the identified set is fully parametrized once an origin is chosen. The parametrization in Q is not linear. Then, in general, the identified set is neither convex, nor even star-convex.

3.2 Examples

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

When $K = 1$, $A = \beta_1 \gamma_1'$, $Q = (q_{11}) = (1)$, the non-negative 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 non-negative matrix A (resp. A') has a unique eigenspace of dimension 1 generated by a non-negative eigenvector, here β_1 (resp. γ_1)⁵.

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 the true matrix", the inequality restrictions in Proposition 1 become :

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

It is easy to check that these inequalities imply $1 - q_{12}q_{21} > 0$. The latter inequality is satisfied, in particular, when we consider local identification in a neighbourhood of

⁵It is easy to check that β_1 (resp. γ_1) is also an eigenvector of AA' (resp. $A'A$). Therefore they are both elements of a singular value decomposition (SVD). Such a model is commonly applied in the network literature with an adjacency matrix, whose leading left and right eigenvectors of the SVD are used to define the "so-called" hub and authority centralities, respectively [see Cai et al. (2021) and the references therein].

$B = (\beta_1, \beta_2), C = (\gamma_1, \gamma_2)$, 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

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

From Proposition 1, we deduce the following result :

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

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

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 [Brie (2015)].

Corollary 1 : Under Assumption A.3 and for $K = \text{Rk}_+(A) = 2$, 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 $\gamma_{1,j_4} > 0, \gamma_{2,j_4} = 0$.

Let us now discuss the degree of under-identification. Since the identification issue of factorial directions up to multiplicative non-negative scalars is solved in representation (2.8), we can focus our attention on the identification of $\pi_k, \beta_k^*, \gamma_k^*, k = 1, 2$. For $K=2$, the decomposition (2.8) 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 derive the closed-form parametrization of the identified set for the given

origin $(\beta_1^*, \beta_2^*), (\gamma_1^*, \gamma_2^*)$ as (see online Appendix 1):

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

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

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

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

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

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 under-identification equal to 2.

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, which are the off-diagonal elements of matrix Q . Therefore, the identified set is a manifold of fixed dimension. These parameters satisfy $K(n+m)$ inequality restrictions by Proposition 1. The degree of under-identification increases rather quickly with the non-negative rank and the subset of active restrictions cannot be derived analytically. To determine these restrictions and provide a simplified definition of the identified set, numerical algorithms are needed, such as the Active Set Sequential Quadratic Programming (SQP) algorithm, some of them being available from Artelys Knitro [see Gill et al. (2002), Liu (2005)] at www.artelys.com.

A similar problem arises in sharp set identification for discrete choice models and treatment effects analysis. The main difference is that those inequality restrictions are linear, e.g. as for $K = 2$ in Section 3.2.2, 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, the identified set is neither convex, nor star-convex, in general.

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, either pointwise, or from a confidence set⁶. These two approaches are linked. For example, the set of all 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 (i.e. an origin) and next deduce all the remaining ones from the origin. This will provide an estimator of the identified set. Alternatively, we can derive directly an asymptotic confidence set of the identified set at a given confidence level by inverting appropriately a test procedure.

In this respect, the identified set described in Proposition 1 has the general form⁷ considered in Shi and Shum (2015), eq. (1.1)-(1.2). However, their results can only be used if the origin is well-defined and has a consistent and asymptotically normally distributed estimator.

Our proposed method of estimation of the identified set of NMFs is outlined as follows: 1) We show the convergence of the set of maximum likelihood (ML) 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 and converges to the identified set, but not necessarily pointwise to a given element of this set; 3) Next, we fix an origin in the interior of the identified set, obtained as the solution of an auxiliary optimization in the spirit of Optimal Control Theory [McCann (1995)]. The objective function of that auxiliary optimization is optimized with respect to an alternative parametrization in $(a, \pi, \beta^*, \gamma^*)$. This step provides a consistent estimator of an origin in the identified set.

The new Identifying Maximum Likelihood (IML) approach is introduced to estimate the identified set, analyze the properties of this set estimator and derive its asymptotic distributional properties. We study analytically the properties of the distribution of the identified set estimator by using the random set theory [Molchanov

⁶We distinguish between the pointwise estimated identified set and the confidence set of the identified set at a given confidence level.

⁷up to the introduction of nuisance slackness parameters [Shi and Shum (2015), Remark p. 497].

and Molinari (2018)]. In addition, we derive the ML estimator of matrix A under a given non-negative rank constraint and provide the efficiency bound of this identifiable matrix-parameter estimator. The approach is feasible because of the properties of the maximum likelihood estimator reviewed briefly below.

4.1 The ML approach

In our framework of nonlinear models with partial identification, the ML estimator(s) does not have a closed-form expression. Hence, we have to distinguish the properties of the statistical convergence of ML estimators to the true identified set when the number T of observations tends to infinity from the numerical convergence of the algorithm used to approximate the ML estimator when the number of iterations tends to infinity⁸.

4.1.1 Consistency of ML Estimators

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

$$l(y_t|y_{t-1}; A_0) = l(y_t|y_{t-1}; B_0C_0'), \quad (4.1)$$

with a non-negative rank K_0 of matrix A_0 assumed to be known. We consider a constrained ML maximization, providing:

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

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

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 NMF_0 of NMF associated with $A_0 = B_0C_0'$, when T tends to infinity.

⁸In this respect, our analysis differs from the literature on exact NMF that implicitly assumes that A_0 is observed and examines the computational complexity, i.e. the possibility of finding a decomposition of A_0 in polynomial time, called the NP-hardness [see Gillis (2020) for a survey]. This literature disregards the uncertainty of observations.

More precisely, let $d(.,.)$ denote the Euclidean distance on $\mathbb{R}^{K(n+m)}$, $NMF_0 = \{(B, C), B \geq 0, C \geq 0, \text{ with } BC' = B_0 C_0'\}$, 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, (\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 of matrix $A = BC'$ does not always have a closed-form. In practice, it is computed numerically from an algorithm, such as a Newton-Raphson algorithm. In our framework of partial identification, a Newton-Raphson algorithm cannot be used jointly for B and C . The reason is that each iteration requires an inversion of the Hessian matrix, which is not invertible due to the identification issue. The AML algorithm, also called the block mirror descent (BMD) algorithm [Hien and Gillis (2021)], or zig-zag algorithm [Hautsch et al. (2023)] solves this issue. In the presence of a multiplicity of NMFs, we apply an AML algorithm [see, e.g. Gourieroux, Monfort and Renault (1990), Kim and Park (2008), Hastie et al. (2015), Gillis (2020), Chapter 8 for alternating least squares, Hien and Gillis (2021) for NMF]. 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 AML algorithm, where at step p , $\hat{B}_{p,T}, \hat{C}_{p,T}$ is computed and then $\hat{B}_{p+1,T}, \hat{C}_{p+1,T}$ are recursively obtained as follows:

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

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

By construction, this AML 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 AML 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 of B and C . However, there is a unique

sequence of AML estimators for given starting values, even though the AML algorithm does not necessarily numerically converge pointwise, due to the identification issue. Moreover, even if it converges numerically to a global maximum⁹, the limit can depend on the starting value and does not necessarily correspond to a point in the interior of the identified set. The dependence on the initialization has been observed in practice and led to a literature on the optimal choice of the starting value [see the discussion in Gillis (2020), Chapter 4 and Esposito (2021)]. This shows a confusion between the effect of non-identification and the fact that the algorithm can stop at a local maximum, instead of global one in some cases.

The discussion of starting values is different when the focus is on the estimation of matrix A under the NMF restriction. Indeed, the matrix A is identifiable even if the pair of matrices B and C is not. For this identifiable parameter matrix A it is possible to prove that the AML estimator and the infeasible ML estimator have identical asymptotic distributions, if the starting value of the AML algorithm is a consistent estimator of A [see Vrahalis et al. (2003), Hautsch et al. (2023)]. Such starting values could be derived from an unconstrained estimator of A if the number of elements in A is not too large, compared to the number of observations.

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

As pointed out in Section 3.2.1, the NMF is essentially unique for $K_0 = 1$. The assumption $K_0 = 1$ greatly simplifies the estimation and explains its frequent use in applied econometrics [see e.g. Cai et al. (2021)]. 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 $\beta'e = \gamma'e = 1$. As an illustration, we consider the dynamic model described in Section 2.2.3, which is the latent parametric model $l(y; \theta)$, where θ is replaced by Ay_{t-1} (see, Section 2.2.3). We get :

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

The partial derivatives of the log-likelihood with respect to a, β, γ are easily derived

⁹See, Grippo and Sciandrossa (2000) for the numerical convergence conditions.

from the partial derivatives of the latent log-likelihood with respect to θ . We have :

$$\left\{ \begin{array}{l} \frac{\partial \log l}{\partial \beta}(y_t|y_{t-1}; A) = a\gamma'y_{t-1} \frac{\partial \log l}{\partial \theta}(y_t; a\beta\gamma'y_{t-1}), \\ \frac{\partial \log l}{\partial \gamma}(y_t|y_{t-1}; A) = ay_{t-1}\beta' \frac{\partial \log l}{\partial \theta}(y_t; a\beta\gamma'y_{t-1}), \\ \frac{\partial \log l}{\partial a}(y_t|y_{t-1}; A) = \gamma'y_{t-1}\beta' \frac{\partial \log l}{\partial \theta}(y_t; a\beta\gamma'y_{t-1}). \end{array} \right. \quad (4.6)$$

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 strict positivity assumption below.

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

Under this strict positivity assumption, the non-negativity-constrained ML estimators of a, β, γ , 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 :

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

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

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

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

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 and Monfort \(1995\)](#), Section 10.3] and its estimate can be computed by standard software.

Remark 3: 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 \(1963\)](#), [Tipping and Bishop \(1999\)](#)]. However, for $Rk(A_0) = Rk_+(A_0) = 1$, the standard asymptotic properties of the SVD estimation method are not valid, as they account neither for the non-negativity of the data, nor 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 by the Perron, Froebenius Theorem, except the eigenvector β_1 of AA' , all other eigenvectors β_2, β_3, \dots 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, i.e. an origin in the identified set, to deduce the sharp identified set. The problem with applying the AML method is that the convergence of the AML estimator to the identified set does not imply its pointwise convergence to a given NMF, which could be used as an origin. Moreover, since the origin is used to derive the asymptotic distribution of the estimator of the identified set, it has to be in the interior of the identified set. This section shows how an IML algorithm can solve these issues, allowing us to derive the asymptotic distribution of the estimated set of admissible NMF's. For expository purpose, we provide in the text only the most relevant assumptions. 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 AML approximation of 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 :

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

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

$$\mathcal{A}_0 = \{a, \pi_k, \beta_k^*, \gamma_k^*, k = 1, \dots, K, \text{ such that } a \sum_{k=1}^K \pi_k \beta_k^* \gamma_k^{*'} = A_0\}.$$

where A_0 is the true value of A and \mathcal{A}_0 denotes the identified set under the normalized parametrization $\alpha = (a, \pi_k, \beta_k^*, \gamma_k^*, k = 1, \dots, K)$.

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, if $l(y_t|y_{t-1}, A) = l(y_t, Ay_{t-1})$ and the components of y_t are non-negative, 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 then also under the alternative parametrization $(a, \pi_k, \beta_k^*, \gamma_k^*, k = 1, \dots, K)$. Therefore, the AML estimator is a function of the underlying (normalized) log-likelihood $\frac{1}{T} L_T(A) = \frac{1}{T} \sum_{t=1}^T \log l(y_t; Ay_{t-1})$, and of the initial (starting) values of the algorithm (and of the number of iterations p). Let the selected initial (starting) value be denoted by α^o . The AML estimator at iteration p can be written as :

$$\hat{\alpha}_T(\alpha^o, p) \equiv \delta\left(\frac{1}{T} L_T(\cdot); \alpha^o, p\right), \quad (4.8)$$

where δ is a deterministic function. Then, for large T , the AML estimator converges asymptotically to the value :

$$\alpha_\infty(\alpha^o, p) = \delta[E_0 \log l(Y_t; A_0 Y_{t-1}); \alpha^o, p], \quad (4.9)$$

that belongs in the set \mathcal{A}_0 if p is large. This limiting value can depend on the starting value α^o of 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 as a direct consequence of the numerical consistency of the AML approximation (for $p \rightarrow \infty$ and T fixed) and of the uniform convergence in Proposition 3 :

Proposition 4: For large T , there exist a function $c(\cdot)$ and a number of iterations p_T such that, for any $p \geq p_T$: $\mathcal{D}[\alpha_T(\alpha^\circ, p), \mathcal{A}_o] < c(\alpha^\circ)/T$, where \mathcal{D} measures the distance between $\alpha_T(\alpha^\circ, p)$ and the set \mathcal{A}_o .

In practice, we can apply the AML algorithm with a given starting value α° and a finite 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 become negligible and have no impact on the asymptotic distribution of the IML estimator derived later. 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 AML with a fixed number of iterations in each intermediate optimization step will also increase the objective function, which 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 $\alpha_\infty(\alpha^\circ; p)$ and of the elements of a matrix Q with unitary diagonal elements, such that :

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

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

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

$$\mathcal{A}_0 = \{\alpha : \alpha = \zeta[\alpha_\infty(\alpha^\circ, p), Q], Q \in \mathcal{Q}[\alpha_\infty(\alpha^\circ, p)]\}, \quad (4.10)$$

where ζ is a known function. Then, the set \mathcal{A}_0 is consistently estimated as:

$$\hat{\mathcal{A}}_T = \{\alpha : \alpha = \zeta[\hat{\alpha}_T(\alpha^\circ, p), Q], Q \in \mathcal{Q}[\hat{\alpha}_T(\alpha^\circ, p)] \equiv \hat{\mathcal{Q}}_T(\alpha^\circ, p)\}. \quad (4.11)$$

The estimation method presented above will allow us to approximate the identified set for a sufficiently large p . However, because of the lack of identification, we do

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

4.3.2 Identifying Maximum Likelihood: Introducing an identification restriction

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 (this is the "global" reduced form reparametrization considered in Shi and Shum (2015) and Chen et al. (2018), Section 5.1.1). These approaches are not suitable in our framework, where the parametrization of the identified set depends on a selected origin in \mathcal{A}_0 . Hence, we introduce indirectly $K(K - 1)$ identification restrictions.

Let us consider an alternative parametrization method. For a given origin, 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. Let $\tilde{\alpha}$ denote the solution of the AML algorithm at step p . $\tilde{\alpha}$ is not necessarily an element in the interior of the identified set. Moreover, it depends on the starting value.

Let us now determine an interior origin α_0^* . We define the extended parameter vector $(q, \tilde{\alpha}) \equiv \alpha$. Using this notation, the NMF $\tilde{\alpha}$ is now represented as $(0, \tilde{\alpha})$. We define the interior origin of the identified set as the optimizer of an additional criterion with respect to the additional parameter q , in the spirit of the Optimal Transport Theory [McCann (1995)]. Several criteria $\tilde{g}(q, \tilde{\alpha}) = g(\alpha)$, where $\alpha \in \mathcal{A}_0$, arise naturally :

i) The concentration of a discrete probability distribution, which is usually measured by $\sum_{k=1}^K (\pi_k \log \pi_k)$. This quantity is negative, attains its minimum for $\pi_k = 1/K, \forall k$, that is a uniform distribution, and increases to zero with the concentration of the distribution. Therefore, given $\tilde{\alpha}$, the factorization provides us the least concentrated latent heterogeneity defined by :

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

where the minimization is in q for a given $\tilde{\alpha}$. This means that the new selected origin is $\alpha_0^* = (q(\tilde{\alpha}), \tilde{\alpha})$, where:

$$q^*(\tilde{\alpha}) = \mathit{Argmax}_{q \in Q(\tilde{\alpha})} - \sum_{k=1}^K \pi_k(q, \tilde{\alpha}) \log \pi_k(q, \tilde{\alpha}),$$

and the domain of q can depend on $\tilde{\alpha}$.

ii) The volume/collinearity criterion is: $g(\alpha) = \det(B^{*'}B^*)$, where $B^* = (\beta_1^*, \dots, \beta_K^*)$, or $g(\alpha) = \det(C^{*'}C^*)$, where $C^* = (\gamma_1^*, \dots, \gamma_K^*)$. As above B^*, C^* depend on $(q, \tilde{\alpha})$, but we do not account for it to keep the notation simple. The above criterion measures the volume of the parallelepiped generated by the columns of B^* (resp. C^*). The larger this volume, the less "collinear" the columns of B^* (resp. C^*)¹⁰. Alternatively, the collinearity measure could be applied to the matrix B itself. Let \tilde{B} denote the corresponding element of $\tilde{\alpha}$. Then we have:

$$B(q, \tilde{\alpha}) = \tilde{B}Q, \text{ and } \det[B(q, \tilde{\alpha})'B(q, \tilde{\alpha})] = \det[Q'\tilde{B}'\tilde{B}Q] = (\det(Q))^2 \det(\tilde{B}'\tilde{B}).$$

The maximization of the collinearity measure is equivalent to the maximization of $(\det(Q))^2$ on the domain $Q(\tilde{\alpha})$. This objective function is independent of $\tilde{\alpha}$, but the constraints imposed on q depend on $\tilde{\alpha}$.

The patterns of these criteria and their combinations are illustrated and discussed in Online Appendix 3 for the example of Section 6. When they are used, the optimum of function g can still be reached on the boundary of the domain of values of $q_{1,2}, q_{2,1}$ rather than in its interior.

iii) The repulsion criterion. It follows from Proposition 1 that the points in the interior of the identified set are such that the elements $\beta_{i,j}^*$ and $\gamma_{i,j}^*$ are strictly positive. Therefore, a point in the interior can be reached by combining the previous criteria with a third one such that $\sum_i \sum_j (\ln(\beta_{i,j}^*) + \ln(\gamma_{i,j}^*))$ that creates a repulsion effect on the boundary of the domain.

To approximate the interior origin α_0^* , an additional optimization step needs to be included in the AML algorithm¹¹. Let the recursive system in this algorithm be

¹⁰These criteria are the analogues of the SVD identification restrictions: $B'B = C'C = Id$, where all factorial directions are orthonormal.

¹¹This additional step is the analogue of step 2 in the estimation approach introduced in Davezies et al. (2025), Section 3.1., where it is applied to an approximate identified set instead of the identified set itself.

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 $\alpha^{(0)}$. The parameter value needs to be in the parameter set. In particular, the associated starting values $B^{(0)}, C^{(0)}$ have to be non-negative and of a given rank K . This rank condition on the starting value is important to ensure non-degenerate behavior of the algorithm at the next steps.

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 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 function $q(\cdot)$ does not depend on the observations.

iii) Find $\alpha^{(p+1)}$ by transforming $\tilde{\alpha}^{(p+1)}$ with the linear transformation $Q^{(p+1)}$ associated with $q^{(p+1)}$ to get $\pi_k^{(p+1)}$, $k = 1, \dots, K$ arranged in a decreasing order¹². More precisely:

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

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 when T, p_T tend to infinity according to Proposition 4. The AML converges to the set \mathcal{A}_0 , but the convergence is not pointwise. The IML converges to a given interior origin 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 and 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

¹²To solve the up-to-permutation identification issue.

largely solved this problem [see e.g. Gill et al. (2002) and Liu (2005) for a proof of numerical convergence]. 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 the AML algorithm.

4.3.3 Asymptotic distributions

The standard asymptotic arguments can be used to derive the asymptotic normality of the IML estimator adjusted to reach α_0^* , since α_0^* is in the interior of \mathcal{A}_0 if the associated π_{0k}^* , $k = 1, \dots, K$ are all distinct¹³. To clarify the role of the intermediate optimization in the IML algorithm, let us first consider a standard information matrix.

In fact, two information matrices arise naturally¹⁴ :

- i) The information based on unconstrained A , is $E_0 \left[-\frac{\partial^2 \log l(y_t|y_{t-1}; A)}{\partial \text{vec } A \partial \text{vec } A'} \right]$. This matrix is invertible by the assumption of identifiable A , but can be of a high dimension.
- ii) The information matrix corresponding to parameter α : $J_0 = E_0 \left[-\frac{\partial^2 \log l(y_t|y_{t-1}; \alpha)}{\partial \alpha \partial \alpha'} \right]$ is constrained by the unit mass restrictions on $\pi, \beta_k^*, \gamma_k^*$, $k = 1, \dots, K$. This matrix has a smaller dimension, but is not of full rank because of the lack of identification.

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

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

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 :

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

¹³This additional condition is analogous to the condition of distinct eigenvalues in the joint spectral decomposition of AA' and $A'A$ in the SVD.

¹⁴For expository purpose, we keep the same notation l for the conditional likelihood as a function of A , or a function of α .

as the additional asymptotic linear restrictions $D'_2\sqrt{T}(\hat{\alpha}_T - \alpha_0^*) = 0$ on $\sqrt{T}(\hat{\alpha}_T - \alpha_0^*)$, say, to be taken into account for the computation of the asymptotic variance of $\hat{\alpha}_T$. More precisely, we get

Proposition 5 : If α_0^* is in the interior of \mathcal{A}_0 , if the associated π_{0k}^* , $k = 1, \dots, K$, are all distinct, if assumptions A.1-A.4 and the additional regularity assumptions a.1-a.2 are satisfied, then, $\sqrt{T}(\hat{\alpha}_T - \alpha_0^*) \xrightarrow{d} N(0, J_0^{11} J_0 J_0^{11})$, 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 $\pi_k, \beta_k^*, \gamma_k^*$, $k = 1, \dots, K$, and D_2 defining the (asymptotic) linearized restrictions corresponding to the intermediate optimizations in the IML algorithm.

Proof: See Appendix 1.1.

In the expression of the asymptotic variance-covariance matrix, the three main components are 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.

Remark 5: The sequence of optimizations cannot be replaced by a penalty term in the objective function as suggested by the machine learning literature either to ensure the (numerical) convergence of the AML algorithm [see e.g. Kim and Park (2008), Schachtner et al. (2011), Hastie et al. (2014)], or to circumvent the curse of dimensionality. 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 a function of T , ensuring the numerical convergence. It is easy to see why this approach would not provide an estimator converging to an interior origin in the identified set. The asymptotic first-order conditions would become $\frac{\partial \log l_T(y, \alpha)}{\partial \alpha} + \lambda_T \frac{\partial g(\alpha)}{\partial \alpha}$. These FOCs are not aligned with the direction allowing us 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 on the IML estimator of α_0^* determines all the uncertainties on the IML estimator of the 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 deterministic 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 (the δ -method), whenever q is not on the boundary of $Q(\alpha_0^*)$. If q is on the boundary, its asymptotic distribution will become a truncated normal, and can be easily found by simulation.

The IML approach can be used to derive the lower and upper bounds on partially identified scalar parameters, and to obtain measures of uncertainty on these bounds. The examples are the minimum and maximum values of functions $\sum_{k=1}^K (\pi_k \log \pi_k)$, $\det(\tilde{B}'\tilde{B})$, $\det(\tilde{C}'\tilde{C})$. Suppose, 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 q_{12}, q_{21} when $K = 2$, for example. The joint asymptotic distribution of these bounds cannot be Gaussian due to the infimum appearing in the formulas of Proposition 2. Note that the joint asymptotic distribution of these two bounds is easily derived by simulations and, by construction, the estimated bounds cannot cross. The IML approach can also be used to derive a confidence set for \mathcal{A}_0 of a desired asymptotic level, following Shi and Shum (2015). However this asymptotic confidence set is difficult to represent graphically due to the high dimension of \mathcal{A}_0 .

4.3.4 Asymptotic distribution of \hat{A}_T

The IML approach helps us find the estimates and confidence intervals of identifiable parameters, such as the elements a_{ij} of matrix A . In practice, even if A_0 is identifiable, we may encounter the curse of dimensionality that complicates 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 (CIs) are likely narrower than the unconstrained ones. The CIs 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} converges to the true value $a_{ij,0}$ that is independent of the choice of interior origin α_0^* . Similarly, its asymptotic variance-covariance matrix is also independent of this choice, i.e. of the selected function \tilde{g} and of the starting values of the IML algorithm. The additional constraints D_2 are introduced only for solving the identification issue.

Proposition 6 The asymptotic distribution of the IML estimator of matrix A does not depend on the interior origin α_0^* in the identified set, i.e. on the choice of the additional optimization criterion.

Proof: See Appendix 1.2.

We have introduced an IML estimator of A_0 under the non-negative rank restriction $Rk_+(A_0) = K_0$, derived its asymptotic Gaussian behavior¹⁵ and found the expression of the associated efficiency bound on this identifiable matrix parameter. 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 $\alpha_0^* = (a_0^*, \pi_{0k}^*, \beta_{0k}^*, \gamma_{0k}^*, k = 1, \dots, K)$. Then, the asymptotic confidence intervals for the identifiable parameter functions of A also share the asymptotic efficiency properties.

The asymptotic expansion and normality of the IML estimator of A could also be used to derive the asymptotic distribution of the log-likelihood ratio test statistic corresponding to the test of hypotheses: $H_0 : Rk_+ A = K - 1$ against $H_0 : Rk_+ A = K$ that is a chi-square distribution with an appropriate degree of freedom. Then, we would deduce a sequence of test procedures to estimate $Rk_+(A)$. Due to the curse of dimensionality we cannot follow a "general to specific" approach, starting from the highest possible rank $K = \inf(n, m)$ and testing it against the alternative of a lower value of K . Instead, we have to start from the smallest value of $K = 1$ and test it against higher values of $K = 2, 3, \dots$

5 Illustration

We consider in this section observations $y_t, t = 1, \dots, T$, on a process satisfying a multivariate Autoregressive Conditional Poisson (ACP) model:

$$y_{it}|y_{t-1} \sim \mathcal{P}(a_i y_{t-1} + \mu_i), \quad a_i \geq 0, \mu_i \geq 0, \quad i = 1, \dots, n, \quad (5.1)$$

¹⁵This asymptotic Gaussian distribution is degenerate because of the reduced rank.

where y_{1t}, \dots, y_{nt} are independent given y_{t-1} . The process is such that $E(Y_t|Y_{t-1}) = AY_{t-1} + \mu$ and ergodic if the eigenvalues of A are of modulus strictly less than 1.

5.1 The true dynamics

We consider a process of dimension four $n = 4$, and the true matrix A_0 of non-negative rank equal to 2. This matrix is given by :

$$\begin{aligned} A_0 &= \frac{1}{96} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} (1, 1, 2, 2) + \frac{1}{96} \begin{pmatrix} 1 \\ 2 \\ 1 \\ 2 \end{pmatrix} (1, 1, 1, 1) \\ &= \frac{1}{96} \begin{pmatrix} 2 & 2 & 3 & 3 \\ 3 & 3 & 4 & 4 \\ 2 & 2 & 3 & 3 \\ 3 & 3 & 4 & 4 \end{pmatrix} = \begin{pmatrix} 0.0208 & 0.0208 & 0.0312 & 0.0312 \\ 0.0312 & 0.0312 & 0.0416 & 0.0416 \\ 0.0208 & 0.0208 & 0.0312 & 0.0312 \\ 0.0312 & 0.0312 & 0.0416 & 0.0416 \end{pmatrix}. \end{aligned} \quad (5.2)$$

and the true intercept is $\mu_0 = (2, 2, 2, 2)'$. The matrix A_0 can be rewritten to highlight its interpretations in terms of probability distributions (see eq. 2.8). We have :

$$A_0 = a_0[\pi_0\beta_{01}^*\gamma_{01}^{*'} + (1 - \pi_0)\beta_{02}^*\gamma_{02}^{*'}], \quad (5.3)$$

where $a_0 = 1/2, \pi_0 = 1/2$,

$$\beta_{01}^* = \frac{1}{4} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \gamma_{01}^{*'} = \frac{1}{6} \begin{pmatrix} 1 \\ 1 \\ 2 \\ 2 \end{pmatrix}, \beta_{02}^* = \frac{1}{6} \begin{pmatrix} 1 \\ 2 \\ 1 \\ 2 \end{pmatrix}, \gamma_{02}^{*'} = \frac{1}{4} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}. \quad (5.4)$$

By Proposition 2, we get :

$$-1 \leq q_{12} \leq 1, \quad -1/2 \leq q_{21} \leq 1/2. \quad (5.5)$$

Therefore the identified set for the matrix :

$$B = B_0Q = \frac{1}{96} \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & q_{12} \\ q_{21} & 1 \end{pmatrix},$$

is convex and generated by the four extreme matrices corresponding to :

$$(q_{12} = -1, q_{21} = -1/2), \quad (q_{12} = -1, q_{21} = 1/2), \quad (q_{12} = 1, q_{21} = -1/2), \quad (q_{12} = 1, q_{21} = 1/2)$$

For instance the first extreme point is :

$$B = \frac{1}{96} \begin{pmatrix} 1/2 & 0 \\ 0 & 1 \\ 1/2 & 0 \\ 0 & 1 \end{pmatrix},$$

and has some zero entries, i.e. satisfies the sparsity property.

We generate simulated trajectories of a four-dimensional count process for $t = 1, \dots, T = 500$ and display in Figure 1 the first 200 values. The initial values are set equal to $y_{it} = 3, i = 1, \dots, 4$ at $t = -50$. The observed values vary between 0 and 10.

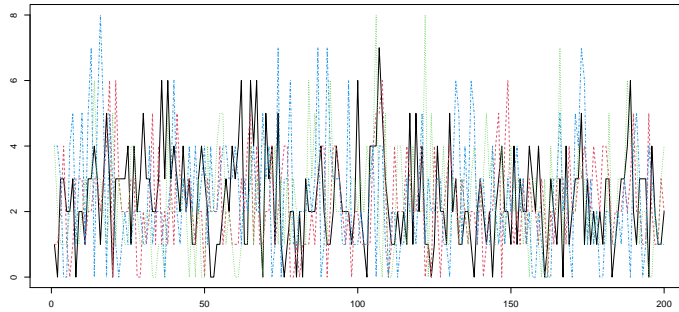


Figure 1: Evolution of y_t

By construction, the effect of lagged values is captured by the sufficient statistics: $z_{1,t} = y_{1,t} + y_{2,t}$ and $z_{2,t} = y_{3,t} + y_{4,t}$ displayed in Figure 2.

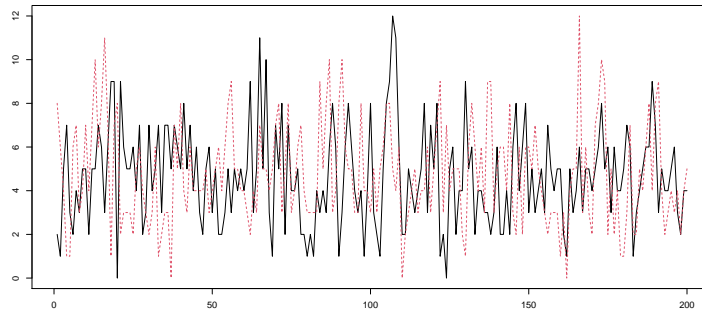


Figure 2: Evolution of the Sufficient Statistics

The process $z_t = (z_{1,t}, z_{2,t})'$ follows a conditional Poisson model with a NMF of non-negative rank equal to 2, i.e. the dimension of z . We observe clustering of small and large values in Figures 1 and 2 because of serial dependence of both processes. The true parameter values chosen in this experiment can complicate the inference because matrix A_0 has two non-negative singular values, the first one 37.9 times bigger than the second one, making it difficult to distinguish between the ranks $\text{Rk}(A)=1$ and $\text{Rk}(A)=2$ in practice. Moreover, the uniform distributed heterogeneity in matrix A_0 factorization complicates the identification.

5.2 Inference

To find the NMF, we consider the maximum likelihood estimation of matrix A from the simulated data. The log-likelihood computed from the component series $i = 1, \dots, 4$ at times $t = 1, \dots, T$ and concentrated with respect to μ is:

$$l(A; y) \approx \sum_{i=1}^4 \sum_{t=2}^T \{ -(\bar{y}_i + a_i(y_{t-1} - \bar{y})) + y_{i,t} \ln(\bar{y}_i + a_i(y_{t-1} - \bar{y})) \} \quad (5.6)$$

where \bar{y} denotes the marginal mean of the series and \approx stands for the equality up to an additive term independent of parameter A . This concentrated log-likelihood function is maximized by applying the IML algorithm in which the additional maximization criterion combines the concentration measure $\pi \ln \pi + (1 - \pi) \ln(1 - \pi)$, the measure of collinearity of the standardized factors β^* , i.e. $|\det(B^* B^*)|$, and the repulsion criterion. The joint use of the three criteria is needed in our experiment. The use of the concentration criterion only in the auxiliary optimization is insufficient to define an origin in the true identified set. It allows us only to reduce the degree of under identification by one. With the additional collinearity criterion, the solution α^* is located on the border of the identified set. Then, the repulsion criterion allows us to reach a point α_0^* in the interior. A discussion of the patterns of these additional objective criteria as a function of Q is provided in the Online Appendix 3.1 for the non-negative rank $K = 2$.

The IML estimated components $q_{1,2}$ and $q_{2,1}$ of matrix Q , based on AML matrices \hat{B}_T and \hat{C}_T are reported in Online Appendix 3.3 and obtained by maximizing the auxiliary objective function with the volume/collinearity and repulsion criteria are

$\hat{q}_{1,2} = 0.1833$ and $\hat{q}_{2,1} = 0.0872$ for $T = 500$ and $\hat{q}_{1,2} = 0.1557$ and $\hat{q}_{2,1} = -0.0393$ for $T = 1000$.

For the sample size $T=500$, we get the IML estimated matrices:

$$\hat{B}_{500}^{IML} = \hat{B}_{500}^{AML} * \hat{Q}_{500} = \begin{pmatrix} 0.0291 & 0.1181 \\ 0.0395 & 0.0939 \\ 0.1331 & 0.0279 \\ 0.1182 & 0.0511 \end{pmatrix}, \quad \hat{C}_{500}^{IML} = \hat{C}_{500}^{AML} * (\hat{Q}'_{500})^{-1} =$$

$$\begin{pmatrix} 0.0926 & 0.3922 \\ 0.1250 & 0.2903 \\ 0.17648 & 0.2112 \\ 0.4542 & 0.0840 \end{pmatrix}, \quad \hat{A}_{500} = \begin{pmatrix} 0.0490 & 0.0379 & 0.0301 & 0.0232 \\ 0.0405 & 0.0322 & 0.0268 & 0.0259 \\ 0.0233 & 0.0248 & 0.0294 & 0.0628 \\ 0.0310 & 0.0297 & 0.0317 & 0.0580 \end{pmatrix},$$

with $\widehat{dist} = 0.5136$, $\hat{\pi} = 0.4881$, and the normalized matrices are:

$$\hat{\beta}_{500}^* = \begin{pmatrix} 0.0911 & 0.4056 \\ 0.1236 & 0.3226 \\ 0.4158 & 0.0959 \\ 0.3693 & 0.1757 \end{pmatrix}, \quad \hat{\gamma}_{500}^* = \begin{pmatrix} 0.1091 & 0.4010 \\ 0.1474 & 0.2969 \\ 0.2080 & 0.2159 \\ 0.5354 & 0.0859 \end{pmatrix},$$

For the sample size $T=1000$, we get the estimated matrices:

$$\hat{B}_{1000}^{IML} = \hat{B}_{1000}^{AML} * \hat{Q}_{1000} = \begin{pmatrix} 0.0480 & 0.0687 \\ 0.0628 & 0.0750 \\ 0.1379 & 0.0220 \\ 0.1225 & 0.0409 \end{pmatrix}, \quad \hat{C}_{1000}^{IML} = \hat{C}_{1000}^{AML} * (\hat{Q}'_{1000})^{-1} =$$

$$\begin{pmatrix} 0.1140 & 0.3718 \\ 0.0794 & 0.3321 \\ 0.1558 & 0.1774 \\ 0.4670 & 0.1901 \end{pmatrix}, \quad \hat{A}_{1000} = \begin{pmatrix} 0.0310 & 0.0266 & 0.0197 & 0.0355 \\ 0.0351 & 0.0299 & 0.0231 & 0.0436 \\ 0.0240 & 0.0183 & 0.0254 & 0.0687 \\ 0.0292 & 0.0233 & 0.0264 & 0.0650 \end{pmatrix},$$

with $\widehat{dist} = 0.4119$, $\hat{\pi} = 0.5777$ and the normalized matrices are:

$$\hat{\beta}_{1000}^* = \begin{pmatrix} 0.1292 & 0.3323 \\ 0.1692 & 0.3630 \\ 0.3715 & 0.1067 \\ 0.3299 & 0.1978 \end{pmatrix}, \quad \hat{\gamma}_{1000}^* = \begin{pmatrix} 0.1397 & 0.3470 \\ 0.0973 & 0.3099 \\ 0.1908 & 0.1655 \\ 0.5720 & 0.1774 \end{pmatrix},$$

where \widehat{dist} is a weighted measure of distance between \hat{A}_T and A_0 (see online Appendix 3.2).

Additional estimation results are provided in on-line Appendix 3. In on-line Appendix 3.2 we show the unconstrained OLS estimation of matrix A based on the Seemingly Unrelated Regressions (SUR) model. The approach is simple to implement, although much less accurate than the IML method because the constraint of

a known rank is not used and the OLS provides negative values of some estimated coefficients, which are incompatible with the parameter set. The results of the AML step of our procedure are given in on-line Appendix 3.3. We observe that the AML estimators of matrix A converge, while the AML estimators of B^* and C^* show much more variation due to the identification issue.

6 Concluding Remarks

We introduce the Identifying Maximum Likelihood (IML) method for estimation of the identified 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 and derive its asymptotic distribution.

Our approach can be used in a variety of applications with a lack of local identifiability. This arises when an element of the identified set is characterized as a solution of auxiliary optimizations, and the identified set can be parametrized given this element considered as a new origin of the identified set (manifold). For example, the approach can be applied to the identification of finite mixtures (see Online Appendix 5). In practice, the dimension of the parametric identified set can be very large and not representable 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 graphically.

REFERENCES

- Anderson, T., (1963) : "Asymptotic Theory for Principal Component Analysis", Ann. Math. Statist., 34, 122-148.
- 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.
- Berman, A., and R., Plemmons (1994) : "Nonnegative Matrices in the Mathematical Sciences", Philadelphia, Classics in Applied Mathematics, SIAM, Elsevier.

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.

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

Cai, J., Yang, D. Zhu, W. Shen, H., and L. Zhao (2021): "Network Regressions and Supervised Centrality Estimation", DP. University of Pennsylvania.

Cameron, C., and P., Trivedi (2008): "Regression Analysis of Count Data", Econometric Society Monographs, 30, Cambridge University Press.

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

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

Davezies, L., D'Haultfoeuille, X., and L., Laage (2025) : "Identification and Estimation of Average Marginal Effects in Fixed Effects Logit Models", WP 2025-02, CREST

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

Engle, R., and J., Rangel (2008): "The Spline GARCH Model for Unconditional Volatility and its Global Macroeconomic Causes", Review of Financial Studies, 21, 1187-1222.

Esposito, F. (2021): "A Review of Initialization Methods for Nonnegative Matrix Factorization: Towards Omics Data Experiments", Mathematics, 9, 10060.

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

Fokianos, K. (2024): "Multivariate Count Time Series Modelling", Econometrics and Statistics, 31, 100-116.

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

Gillis, N. (2020): "Nonnegative Matrix Factorization", SIAM , Philadelphia.

Gourieroux, C., and Y., Lu (2019): "Negative Binomial Autoregressive Process with Stochastic Intensity", Journal of Time Series Analysis, 40, 225-247.

Gourieroux, C., and Y., Lu (2023): "Susceptible-Infected- Recovered Model with Stochastic Transmission", Canadian Journal of Statistics, forthcoming.

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

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.

Grippo, L., and M. Sciandrone (2002): "On the Convergence of Block Nonlinear Gauss Seidel Method under Convex Constraints", Optimization Research Letters, 26, 127-136.

Hafner, C., and O. Linton (2010): "Efficient Estimation of a Multivariate Multiplicative Volatility Model", Journal of Econometrics, 159, 55-73.

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

Hastie, T., Mazunder, M., Lee, J. and R., Zadeh (2014): "Matrix Completion and Low Rank SVD via Fast Alternating Least Squares", Journal of Machine Learning Research, 16, 3367-3402.

Hautsch, N., Ohrkin, O., and R., Alexamder (2023): "Maximum Likelihood Estimation Under the Zig-Zag Algorithm", Journal of Financial Econometrics, 21, 1346-1375.

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

Hien, L., and N., Gillis (2021): "Algorithm for Nonnegative Matrix Factorization with the Kullblack-Leibler Divergence", Arxiv 2010-01935V2.

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

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

Kim, H., and H., Park (2008) : "Nonnegative Matrix Factorization Based on Alternating Nonnegativity Constrained Least Squares and Active Set Method", SIAM Journal of Matrix Anal. Appl., 30, 716-730.

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.

- Lee, D., and H., Seung (1999) : "Learning the Parts of Objects by Nonnegative Matrix Factorization", *Nature*, 401, 788-791.
- Liu, X. (2005) : "Global Convergence on an Active Set SQP for Inequality Constrained Optimization", *Journal of Computational and Applied Mathematics*, 180, 201-211.
- Lu, Y., Zhang, J. and W. Zhu (2024): "Cyber Risk Modeling: a Discrete Multivariate Count Process Approach", *Scandinavian Actuarial Journal*, 6, 1-31.
- McCann, R. (1995): "Existence and Uniqueness of Monotone Measure Preserving Maps", *Duke Math. J.*, 80, 309-324.
- McCullagh, P., and J. Nelder (1989): "Generalized Linear Models", 2nd ed., London, Chapman and Hall.
- Molchanov, I. and F. Molinari (2018): "Random Sets in Econometrics", Cambridge University Press.
- Paatero, P., and U., Tapper (1994): "Positive Matrix Factorization: A Nonnegative Factor Model with Optimal Utilisation of Error Estimates of Data Values", *Environmetrics*, 5, 111-126.
- 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.
- Shi, X., and M. Shum (2015): "Simple Two-Stage Inference for a Class of Partially Identified Models", *Econometric Theory*, 31, 493-520.
- Tipping, H., and C., Bishop (1999) : "Probabilistic Principal Component Analysis", *JRSS, B*, 61, 611-622.
- Vrahalis, M., Magoulas, G. and V. Piagianakos (2003): "From Linear to Nonlinear Iterative Methods", *Applied Numerical Mathematics*, 45, 59-77.

Appendix 1: Proofs of Propositions 5 and 6

1. Proof of Proposition 5

- i) The proof is standard and based on the asymptotic expansion of the first-order conditions on the Lagrangean 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 :

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

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 score in (a.1) is asymptotically normally distributed with mean zero and variance J .

Then, 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 : $J_0\theta + D\lambda = 0, \quad D'\theta = 0$.

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 : $(Id - P)J_0\theta = 0, \quad D'\theta = 0$, 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 of the above system, if and only if : $Rk((Id - P)J_0) = \dim \alpha - \dim q - 1 - 2K$. This statement is Assumption a.2 viii) in on-line Appendix 2.

2. Proof of Proposition 6

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 (a.1). Asymptotically, a change of the benchmark modifies the matrix D_2 as well as the associated Lagrange multipliers by linear transformations R and R^{-1} , respectively. Then, the first-order conditions provide the same solution for $\sqrt{T}(\hat{\alpha}_T - \alpha_0^*)$ when D_2 is replaced by $\tilde{D}_2 = D_0R$ and $\hat{\lambda}_{2T} - \lambda_{20}^*$ by $\hat{\lambda}_{2T} - \tilde{\lambda}_{20}^* = R^{-1}(\hat{\lambda}_{2T} - \lambda_{20}^*)$, where R is invertible. This proves that the asymptotic variance-covariance matrix is independent of the choice of the additional optimization criterion.

¹⁶We cannot use the usual block formula to compute J_0^{-1} [see e.g. Gourieroux and Monfort (1995), 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 .