Centre de recherche en économie et statistique

Série des Documents de Travail

n° 2015-03

Statistical Inference for Independent Component Analysis

C.GOURIEROUX[1]
A.MONFORT[2]

[1] CREST and University of Toronto

[2] CREST and Banque de France

Statistical Inference for Independent Component Analysis

C., GOURIEROUX [1] and A., MONFORT [2]

June, 2015

The authors gratefully acknowledge support of the chair LCL : "New Challenges for New Data" and of the LABEX : Finance et Croissance Durable.

[1]CREST and University of Toronto.
[2]CREST and Banque de France.

Statistical Inference for Independent Component Analysis

Abstract

The modelling of error terms in multivariate dynamic models by independent component analysis (ICA) is required for reliable impulse response analysis in macroeconomic applications. Since the introduction of ICA by Comon (1994), a large number of semi-parametric estimation methods have been proposed for "orthogonalizing" the error terms. These methods can be pseudo-maximum likelihood (PML) approaches, recursive PML, or moment methods. However several of these approaches are not consistent, and the other ones can be significantly subefficient. The aim of our paper is to derive the asymptotic properties of the PML approaches, in particular to study their consistency (or lack of consistency). Moreover we introduce covariance estimators and explain how to improve their efficiency. Finally we discuss the empirical likelihood approach.

Keywords : Independent Component Analysis, Pseudo-Maximum Likelihood, Method of Moments, Empirical Likelihood, Identification, Cayley Transform.

1 Introduction

Let us consider n observed variables $Y = (y_1, \ldots, y_n)'$, which are linear combinations of n independent unobserved sources $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$:

$$Y = C\varepsilon, \tag{1.1}$$

where the components ε_i are zero-mean, and the matrix C is invertible.

C is called the "mixing matrix" and C^{-1} the "unmixing matrix". The problem of independent component analysis (ICA) [3] is to identify C and ε from the knowledge of Y, or, in other words, to estimate consistently C and the distribution of ε, from a large number of observations $Y_1, \ldots, Y_T$ of vector Y.

[3] In signal processing, the components of ε are called "sources", the components of Y are called "sensors" and the ICA problem "blind separation of sources". Other terminologies are "sources/mixtures", "signal/mixtures", or "multiple input/multiple output" (MIMO).

If ε is Gaussian, the distribution of Y is Gaussian too with zero-mean and a variance-covariance matrix CC'. From the knowledge of the distribution of Y, we can identify CC', but not matrix C itself. For instance, if $C^* = CQ$, where Q is an orthogonal matrix, we have $C^*C^{*'} = CC'$. Thus there is a problem of both local and global identification, since C is identified up to an orthogonal matrix. However the lack of identification almost disappears, if we assume that the components of ε are independent, not Gaussian. The theorem below has been derived in Eriksson, Koivunen (2004) [see also Comon (1994), Th. 11].

Theorem [Eriksson, Koivunen (2004), Th 3] :

Let us consider the independent component model : $Y = C\varepsilon$. Under the following conditions :

i) C is invertible.

ii) The components $\varepsilon_1, \ldots, \varepsilon_n$ are independent, with at most one Gaussian distribution,

then matrix C is identifiable up to the post multiplication by DP, where P is a permutation matrix and D a diagonal matrix with non zero diagonal elements.

In other words C is identifiable up to a permutation of indexes, and signed scaling, $\varepsilon_{i,t} \to \pm\sigma_i\varepsilon_{i,t}, \sigma_i > 0, i = 1, \ldots, n$, say.

Thus, for independent non-Gaussian sources, the only cause of local lack of identification is through the positive scaling. The permutation and change in signs of columns of C create a global lack of identification, but not a local one.

The local identification problem i.e. the possibility of replacing C by CD, where D is a diagonal matrix with strictly positive diagonal elements, can be avoided by introducing identification restrictions. Several sets of identification restrictions (SIR) have been considered in the literature, that are,

SIR 1 : $c_{i,i} = 1, i = 1, \ldots, n$ [see e.g. Jutten, Herault (1991), Comon, Jutten, Herault (1991), eq. (3), Pham, Garat (1997), p1714, Ilmonen, Paindaveine (2015)].

SIR 2 : $c_i'c_i = 1, i = 1, \ldots, n$, where c_i denotes the i^{th} column of matrix

C [see e.g. Comon (1994), Section 5.1, Pham, Garat (1997), p1714].

or similar sets of identification restrictions written on the diagonal elements $c^{i,i}$, or on the rows $c^i, i = 1, \ldots, n$, of the demixing matrix C^{-1} :

SIR* 1: $c^{i,i} = 1, i = 1, \ldots, n$,

SIR* 2 : $c^i c^{i'} = 1, i = 1, \ldots, n$ (implicitly used in the one-unit Fast ICA algorithm, see Sections 3.1, 3.2)

Finally stronger conditions can be introduced as in the following set of restrictions :

SIR 3 : C is an orthogonal matrix : $C'C = Id$ [see e.g. Hyvarinen (1997), eq. 13, Vlassis (2001), eq.23, Hastie, Tibshirani (2002), eq.6].

If the error ε is standardized $V(\varepsilon) = Id$, these restrictions may imply constraints on the distribution of vector Y, such as $V(Y_t) = Id$ for SIR 3. This restriction can be asymptotically satisfied if the data are jointly prewhitened.

Note that the restrictions SIR1 and SIR*1 have a major drawback, since they assume implicitly that all diagonal elements are different from zero. Thus they exclude a priori some noncausal features between the variables and can bias the impulse response analysis in a dynamic model with independent shocks.

Whenever the independent component model is locally identified, we can expect consistent semi-parametric estimation methods based on an i.i.d. sample $Y_1, \ldots, Y_T$. Two types of approaches have been initially proposed in the literature, that are, essentially pseudo-maximum likelihood approaches and moment methods . They differ by the form of the objective function, but also by the set of identification restrictions (SIR 1-SIR3), which is used. These estimation methods have been introduced mainly in the literature on signal processing and data analysis with a focus on the numerical convergence and computational complexity of the algorithm used to get the estimate [see e.g. Common (1994), Sections 4.2, 4.3., Hyvarinen (1997), Section 6, Hyvarinen (1999), Hyvarinen, Oja (2000), Section 6.1]. As noted in Ilmonen et al. (2012), "In the computer science communities ICA procedures are usually

seen as algorithms rather than estimates with their statistical properties." The statistical properties of the estimators, such as their consistency, asymptotic normality, or asymptotic efficiency are rarely considered. This explains why several standard methods for ICA proposed in the literature or in the softwares are not statistically consistent, or, when they are statistically consistent, are poorly efficient.

In the following sections, we carefully analyze the statistical properties of these estimation approaches. We first consider in Section 2 the pseudo maximum likelihood (PML) approaches for estimating matrix C under SIR3. Although they maximize a misspecified log-likelihood function, they provide consistent estimators. Then we derive the asymptotic distribution of these PML estimators. In Section 3, we discuss the other standard PML approaches proposed in the literature. We first show that the one-unit algorithm using identification restrictions as SIR2, SIR*2 provides estimators, which are not statistically consistent. For large dimension n the optimization of the pseudo likelihood under SIR 3 can be numerically cumbersome. We extend the analysis to recursive PML approaches under SIR3, which compute in a recursive way the estimators of the columns of C.

The PML and recursive PML approaches under SIR3 provide consistent estimators which are not necessarily very accurate. The consistency of PML and recursive PML estimators under SIR3 is due to their interpretation as specific covariance estimators. In Section 4 we develop the complete theory of generalized covariance estimators and explain how to improve their efficiency.

The PML, recursive PML and Covariance estimators focus on the semi-parametric estimation $\hat{C}$ of matrix C. Then these estimates can be used to derive approximations of the sources as $\hat{\varepsilon}_t = \hat{C}'Y_t$, and nonparametric functional estimators of the distributions of the sources. Section 5 concludes. The asymptotic results are derived in the Appendices.

2 Pseudo-Maximum Likelihood Approach (under SIR3)

Let us discuss the consistency and the asymptotic properties of pseudo maximum likelihood estimators of matrix C. We first consider the working case of observations such that :

$$Y_t = C_0 \varepsilon_t, \tag{2.1}$$

where $E_0(Y_t) = 0, V_0(Y_t) = Id, E_0(\varepsilon_t) = 0, V_0(\varepsilon_t) = Id$ and the components $\varepsilon_{1,t}, \ldots, \varepsilon_{n,t}$ are assumed both cross-sectionally and serially independent, with unknown true probability density functions (p.d.f.) $f_{i,0}(\varepsilon_i), i = 1, \ldots, n$. In this special framework the C_0 matrix is orthogonal $C_0 C_0' = Id$, which is the set of identification restrictions SIR 3, and is identifiable up to a permutation of index i and changes of sign of its columns, if at most one of the true p.d.f. is Gaussian. [4]

Then we explain how the results of the working case can be extended to a model :

$$Y_t = a(X_t, \theta_0) + \Sigma_0^{1/2} C_0 \varepsilon_t, \tag{2.2}$$

where $E_0(Y_t|X_t) = a(X_t; \theta_0), V_0(Y_t|X_t) = \Sigma_0, E_0(\varepsilon_t) = 0, V_0(\varepsilon_t) = Id.$

2.1 Pseudo-Maximum Likelihood Estimator

Let us introduce a set of p.d.f. $g_i(\varepsilon_i), i = 1, \ldots, n$, and consider the pseudo log-likelihood function :

$$\log l_T(C) = \sum_{t=1}^{T} \sum_{i=1}^{n} \log g_i(c_i' Y_t), \tag{2.3}$$

where c_i is the i^{th} column of matrix C (or c_i' is the i^{th} row of C^{-1}). The log-likelihood function (2.3) is computed as if the errors $\varepsilon_t' s$, were serially independent, with an identical distribution, the p.d.f. of $\varepsilon_{i,t}$ being $\varepsilon_{i,t} \sim g_i(\varepsilon_i)$, noting that $\varepsilon_t = C' Y_t$ and $|\det C| = 1$, since C is orthogonal. Then a pseudo maximum likelihood (PML) estimator of matrix C maximizes the pseudo log-likelihood function taking into account the condition that C is orthogonal. This optimization problem can be written as :

[4] When the sources are cross-sectionally independent, but serially correlated with distinct spectra, they can be identified by second-order methods, that is, from the knowledge of autocovariances only. This possibility to identify by means of the dynamics of the sources is not considered here. It is the basis of second-order estimation methods as AMUSE [Tong et al. (1990)], or SOBI [Belouchrani et al. (1997)], Gaussian PML written in the frequency domain [Pham, Garat (1997), Section 3], or based on canonical correlations [Degerine, Mulki (2000)].

$$\hat{C}_T = \arg\max_C \sum_{t=1}^{T} \sum_{i=1}^{n} \log g_i(c_i' Y_t), \tag{2.4}$$
$$s.t.\ C'C = Id.$$

The optimization problem can also be considered after the elimination of the identification restrictions, that is, after parametrizing the orthogonal matrix C. It is known that any orthogonal matrix with no eigenvalue equal to -1 can be written as :

$$C(A) = (Id + A)(Id - A)^{-1}, \tag{2.5}$$

where A is a skew symmetric (or antisymmetric) matrix, such that $A' = -A$. This is the Cayley's representation of an orthogonal matrix. Moreover, this orthogonal matrix is in a one-to-one relationship with A, since we get :

$$A = (C(A) + Id)^{-1}(C(A) - Id). \tag{2.6}$$

Thus, the PML estimator of matrix C can be alternatively derived as $\hat{C}_T = C(\hat{A}_T)$, where :

$$\hat{A}_T = \arg\max_A \sum_{t=1}^{T} \sum_{i=1}^{n} \log g_i[c_i(A)' Y_t], \tag{2.7}$$

and the optimization is with respect to the parameters characterizing A, that are the subdiagonal elements of A : $a_{i,j}, i > j$.

2.2 The finite sample first-order conditions (FOC)

The FOC can be written either on the constrained optimization problem (2.4), or on its parametrized version (2.7). We give in Appendix 1 the closed form expressions of the derivatives of $C(A)$ with respect to A, which can be used to derive the FOC for the model written under the parametric form. We focus below on the FOC for problem (2.4).

Let us distinguish the different restrictions on matrix C :

$$c_i' c_j = 0, i < j \text{ and } c_i' c_i = 1, i = 1, \ldots, n,$$

and introduce the associated Lagrange multipliers denoted $\lambda_{i,j} = \lambda_{j,i}$, if $i \neq j$, and $\lambda_{i,i}/2$, when both indices are equal.

Then the FOC are :

$$\begin{cases} \sum_{t=1}^{T} Y_t \frac{d \log g_i}{d\varepsilon}(\hat{c}_i' Y_t) - \sum_{j=1}^{n} \hat{\lambda}_{i,j} \hat{c}_j = 0, i = 1, \ldots, n, \\ \\ \hat{c}_i' \hat{c}_j = 0, i < j, \hat{c}_i' \hat{c}_i = 1, i = 1, \ldots, n. \end{cases} \tag{2.8}$$

We get $n^2+n(n-1)/2+n$ conditions for the $n^2+n(n-1)/2+n$ unknowns, that are the $\hat{c}_{i,j}, \hat{\lambda}_{i,j}, i < j$, and $\hat{\lambda}_{i,i}, i, j = 1, \ldots, n$.

Premultiplying the first subsystem of (2.8) by $\hat{C}'$ and taking into account the constraints on the orthogonal matrix $\hat{C}$, we see that the finite sample FOC are equivalent to :

$$\begin{cases} \sum_{t=1}^{T} \hat{c}_j' Y_t \frac{d \log g_i}{d\varepsilon}(\hat{c}_i' Y_t) - \hat{\lambda}_{i,j} = 0, i, j = 1, \ldots, n, \\ \\ \hat{c}_i' \hat{c}_j = 0, i < j, \hat{c}_i' \hat{c}_i = 1, i = 1, \ldots, n. \end{cases}$$

Since $\hat{\lambda}_{i,j} = \hat{\lambda}_{j,i}$, it is possible to derive from this system the equations giving $\hat{C}$. They are :

$$\begin{cases} \sum_{t=1}^{T} \hat{c}_j' Y_t \frac{d \log g_i}{d\varepsilon}(\hat{c}_i' Y_t) - \sum_{t=1}^{T} \hat{c}_i' Y_t \frac{d \log g_j}{d\varepsilon}(\hat{c}_j' Y_t) = 0, i < j, \\ \\ \hat{c}_i' \hat{c}_j = 0, i < j, \hat{c}_i' \hat{c}_i = 1, i = 1, \ldots, n. \end{cases} \tag{2.9}$$

Thus the FOC of the constrained optimization problem (2.4) lead to a subsystem giving the estimate of C.

2.3 Consistency

To derive conditions for the consistency of the PML estimators, we have to consider the associated asymptotic optimization problem and the asymptotic FOC. We have already made the following assumptions on the sources $\varepsilon_t' s$:

Assumption A.1 :
i) The shocks ε_t are i.i.d. with $E_0(\varepsilon_t) = 0, V_0(\varepsilon_t) = Id$.
ii) The components $\varepsilon_{1,t}, \ldots, \varepsilon_{n,t}$ are mutually independent.

In addition we make the following assumption on the p.d.f. of the sources :

Assumption A.2 :
i) The functions $\log g_i, i = 1, \ldots, n$, are twice continuously differentiable.

ii) $sup_{C:C'C=Id}|\sum_{i=1}^{n} \log g_i(c_i'y)| \leq h(y)$, where $E_0[h(Y)] < \infty$.

From Assumption 1 and 2 ii), we know that the finite sample objective function : $Q_T(C) = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n} \log g_i(c_i'Y_t)$ tends almost surely uniformly to the asymptotic one $Q_\infty(C) = E_0[\sum_{i=1}^{n} \log g_i(c_i'Y_t)]$.

Moreover the parameter set, that is, the set of orthogonal matrices is compact. Then the uniform integrability in Assumption A.2 ii) implies the uniform convergence of Q_T towards Q_∞, and the convergence of the optimizers of Q_T to the set of optimiser of Q_∞ [Jennrich (1969), Gourieroux, Monfort (1995), vol 2, chapter 24]. Finally the later optimizers can be analyzed by means of the asymptotic FOC. This approach is followed below.

The asymptotic optimization problem is :

$$\max_C L_\infty(C) = \max_C \text{plim}_{T\to\infty} \frac{1}{T} \log l_T(C) \equiv \max_C \sum_{i=1}^{n} E_0[\log g_i(c_i'Y_t)], \quad (2.10)$$

s.t. $c_i'c_j = 0, i < j, c_i'c_i = 1, i, j = 1, \ldots, n$ with Lagrange multipliers $\lambda_{i,j,0}, \lambda_{i,i,0}/2$. The asymptotic FOC are :

$$\begin{cases} E_0[Y_t \frac{d \log g_i}{d\varepsilon}(c_i'Y_t)] - \sum_{j=1}^{n} \lambda_{i,j} c_j = 0, i = 1, \ldots, n, \\ c_i'c_j = 0, i < j, c_i'c_i = 1, i, j = 1, \ldots, n. \end{cases}$$

By premultiplying the first rows by c_k', by using the conditions of orthogonal matrix and the equality $\lambda_{i,j} = \lambda_{j,i}$, we see that the asymptotic FOC are equivalent to :

$$\begin{cases} \lambda_{i,j} = E_0[c_j' Y_t \dfrac{d \log g_i}{d\varepsilon}(c_i' Y_t)] = E_0[c_i' Y_t \dfrac{d \log g_j}{d\varepsilon}(c_j' Y_t)] = \lambda_{j,i}, i \neq j, \\ \\ \lambda_{i,i} = E_0[c_i' Y_t \dfrac{d \log g_i}{d\varepsilon}(c_i' Y_t)], i = 1, \ldots, n. \end{cases} \tag{2.11}$$

We deduce the following property :

Proposition 1 : The values $C_0, \lambda_{i,j,0} = 0, i < j, \lambda_{i,i,0} = E_0[\varepsilon_{i,t} \dfrac{d \log g_i(\varepsilon_{i,t})}{d\varepsilon}]$, $i = 1, \ldots, n$ are solutions of the asymptotic FOC.

Proof : Indeed replacing the $c_i' s$ by their true values, we get :

$$\lambda_{i,j,0} = E_0[\varepsilon_{j,t} \frac{d \log g_i(\varepsilon_{i,t})}{d\varepsilon}] = E_0[\varepsilon_{i,t} \frac{d \log g_j(\varepsilon_{j,t})}{d\varepsilon}] = \lambda_{j,i,0}.$$

Then, by the independence of $\varepsilon_{i,t}, \varepsilon_{j,t}$ for $i \neq j$, we get :

$$E_0[\varepsilon_{j,t} \frac{d \log g_i(\varepsilon_{i,t})}{d\varepsilon}] = E_0(\varepsilon_{j,t}) E_0[\frac{d \log g_i(\varepsilon_{i,t})}{d\varepsilon}] = 0,$$

since $\varepsilon_{j,t}$ is zero-mean. The conclusion follows.

QED

We deduce a necessary identification assumption for C_0.

Assumption A.3 : Identification from the asymptotic FOC

The only solution of the system of equations :

$$\begin{cases} E_0[c_j' Y_t \dfrac{d \log g_i}{d\varepsilon}(c_i' Y_t)] = E_0[c_i' Y_t \dfrac{d \log g_j}{d\varepsilon}(c_j' Y_t)], i \neq j, \\ \\ C'C = Id, \end{cases}$$

is $C = C_0$.

Assumption A.3 implies restrictions on the choice of the pseudo p.d.f. g_i.

Proposition 2 : If there exist $i, j, i \neq j$, such that the pseudo-distributions g_i and g_j are Gaussian $N(0,1)$, Assumption A.3 is not satisfied.

Proof : Indeed in this case : $\frac{d \log g_i}{d\varepsilon}(c_i' Y_t) = -c_i' Y_t$ and $\frac{d \log g_j}{d\varepsilon}(c_j' Y_t) = -c_j' Y_t$, and the associated $(i.j)$ condition in Assumption A.3 becomes :

$$E_0(c_j' Y_t c_i' Y_t) = E_0(c_i' Y_t c_j' Y_t).$$

This condition is satisfied for any C, not for C_0 only.

QED

Even if Assumption A.3 is satisfied, we are not sure that matrix C_0 corresponds to a maximum of the asymptotic optimization problem. To check this property, we can consider a second-order expansion of $L_\infty(C)$ in a neighbourhood of the true values. It is shown in Appendix A.2.1 that the asymptotic objective function is locally concave under the following assumption :

Assumption A.4 : Local concavity

The asymptotic objective function is locally concave in a neighbourhood of C_0 if and only if,

$$E_0\left[\frac{d^2 \log g_i(\varepsilon_{i,t})}{d\varepsilon^2} + \frac{d^2 \log g_j(\varepsilon_{j,t})}{d\varepsilon^2} - \varepsilon_{j,t}\frac{d \log g_j(\varepsilon_{j,t})}{d\varepsilon} - \varepsilon_{i,t}\frac{d \log g_i(\varepsilon_{i,t})}{d\varepsilon}\right] < 0, \forall i < j.$$

This condition is in particular satisfied under the following set of conditions derived in Hyvarinen (1997), Th1 [see also Hyvarinen, Karhunen, Oja (2001), Th. 8.1]. :

$$E_0[\frac{d^2 \log g_i(\varepsilon_{i,t})}{d\varepsilon^2} - \varepsilon_{i,t}\frac{d \log g_i(\varepsilon_{i,t})}{d\varepsilon}] < 0, i = 1, \ldots, n. \qquad (2.12)$$

However, this set of conditions is sufficient, but not necessary.

Then, we have the following consistency result :

Proposition 3 : Under Assumptions A1-A4 the PML estimator of C exists asymptotically and is a consistent estimator of C_0.

This means that the misspecification of pseudo-distributions g_i has no effect on the consistency of these specific PML estimators. This is easily understood when we consider the asymptotic FOC in (2.11). They simply correspond to zero moment conditions written on :

$$c_j' Y_t \frac{d \log g_i}{d\varepsilon}(c_i' Y_t) - c_i' Y_t \frac{d \log g_j}{d\varepsilon}(c_j' Y_t), i < j.$$

Also note that the consistency result is still valid with a choice of g_i not being a p.d.f., but the interpretation as misspecified ML is more appealing.

2.4 Asymptotic Distribution of the PML Estimator

The asymptotic accuracy of the PML estimator depends on the choice of the pseudo p.d.f. Its asymptotic distribution is derived in Appendix 4.

Proposition 4 : Under Assumptions A1-A4, the PML estimator of C_0 is asymptotically normal, with speed of convergence $1/\sqrt{T}$. Its asymptotic variance-covariance matrix is given in Appendix 4.

For illustration, let us consider the bivariate case $n = 2$. The asymptotic expansion of the FOC shows that :

$$\sqrt{T} \begin{pmatrix} \hat{c}_1 - c_{1,0} \\ \hat{c}_2 - c_{2,0} \end{pmatrix} = \begin{bmatrix} \gamma_{1,2} c_{2,0}' & \gamma_{2,1} c_{1,0}' \\ c_{10}' & c_{20}' \\ c_{10}' & 0 \\ 0 & c_{20}' \end{bmatrix}^{-1} \begin{bmatrix} Z \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

$$\text{where :} \quad \gamma_{i,j} = E_0[\frac{d^2 \log g_i(\varepsilon_{i,t})}{d\varepsilon^2}] - E_0[\varepsilon_{j,t} \frac{d \log g_j(\varepsilon_{j,t})}{d\varepsilon}],$$

$$Z \sim N(0, w^2),$$

$$w^2 = E_0\{[\frac{d \log g_1(\varepsilon_{1,t})}{d\varepsilon}]^2\} + E_0\{[\frac{d \log g_2(\varepsilon_{2,t})}{d\varepsilon}]^2\}$$

$$-2E_0[\varepsilon_{1,t} \frac{d \log g_1(\varepsilon_{1,t})}{d\varepsilon}] E_0[\varepsilon_{2,t} \frac{d \log g_2(\varepsilon_{2,t})}{d\varepsilon}].$$

The expression of the asymptotic variance can be simplified in the bivariate case (see Appendix 5 1)). We get :

$$V_{as}[\sqrt{T}(vec\hat{C} - vecC_0)] = \frac{w^2}{(\gamma_{1,2} + \gamma_{2,1})^2} \begin{pmatrix} c_{2,0}c'_{2,0} & -c_{2,0}c'_{1,0} \\ -c_{1,0}c'_{2,0} & c_{1,0}c'_{1,0} \end{pmatrix}. \quad (2.13)$$

These closed form expressions can facilitate the consistent estimation of the asymptotic variance of $\hat{C}$. Indeed, from the PML estimates $\hat{C}$ we deduce the approximated errors $\hat{\varepsilon}_t = \hat{C}'Y_t$. Therefore $\gamma_{i,j}, w^2$ are easily consistently estimated by replacing the theoretical expectation by its sample counterpart and the errors ε by their approximations $\hat{\varepsilon}$. For instance, we can take :

$$\hat{\gamma}_{i,j} = \frac{1}{T} \sum_{t=1}^{T} \frac{d^2 \log g_i(\hat{\varepsilon}_{i,t})}{d\varepsilon^2} - \frac{1}{T} \sum_{t=1}^{T} [\hat{\varepsilon}_{j,t} \frac{d \log g_i(\hat{\varepsilon}_{j,t})}{d\varepsilon}].$$

For $n = 2$, the elements of C generate a manifold of dimension 1 (see Appendix 5). Thus the asymptotic variance-covariance matrix has rank 1. It has been suggested in Pham, Garat (1997), Section 2.B., to also consider the asymptotic distribution of transformations of $\hat{C}$ such as :[5]

$$\hat{\Delta} = Id - C^{-1}\hat{C} = Id - C'\hat{C}. \quad (2.14)$$

We show in Appendix 5, ii) that :

[5]For expository purpose we have changed their definition of the so-called contamination coefficients initially defined as :

$$\hat{\Delta} = Id - \hat{C}^{-1}C$$

$$V_{as}[\sqrt{T}vec\hat{\Delta}] = \frac{\omega^2}{(\gamma_{1,2}+\gamma_{2,1})^2}\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{2.15}$$

Thus, after this transformation the asymptotic accuracy of $\hat{\Delta}$ no longer depends on matrix C, but only on the distributional properties of the sources and of the pseudo p.d.f.

Finally note that the multiplicative factor function $\omega^2/(\gamma_{1,2}+\gamma_{2,1})^2$ differ from the multiplicative factors derived in Hyvarinen (1997), eq. 15, or in Pham, Garat (1997), where the restrictions on C required for identification do not seem to have been fully taken into account.

The asymptotic accuracy of the PML estimator depends on the choice of the pseudo p.d.f.. Since the ML estimator is asymptotically efficient, we immediately deduce the following Corollary :

Corollary 1 : The asymptotic accuracy of the PML estimator is maximal if $g_i = f_{i,0}$, i.e. if the pseudo p.d.f. is equal to the true p.d.f..

The corollary above raises three comments :

i) The practice of selecting a pseudo p.d.f. as far as possible to a Gaussian distribution, for instance by optimizing a measure of distance to Gaussianity such as the negentropy or an approximation of the negentropy by third and fourth-order cumulants [see e.g. Hyvarinen, Karhunen, Oja (2001), p111, 222-223] is suboptimal, [6] especially, when the true distribution is close to Gaussian.

ii) The asymptotic efficiency for the estimation of parameter C could be reached in two steps by an adaptive estimation approach. In a first step , C is estimated by a non efficient PML approach. The corresponding estimate is used to compute the residuals as : $\hat{\varepsilon}_t = \hat{C}'Y_t, t = 1, \ldots, T$. Next the approximated sources $\hat{\varepsilon}_{i,t}, t = 1, \ldots, T$ can be used to estimate nonparametrically the densities $f_{i,0}, i = 1, \ldots, n$. In a second step the PML approach

[6] See Kaiser (1958) for an early version of such an idea, or the choice $g_i(y) = sech^2(y)/2$, whose associated score function is $2tanh(y)$ introduced in the informax algorithm [Bell, Sejnowski (1995)].

is reapplied with $g_i = \hat{f}_i, i = 1, \ldots, n$, where $\hat{f}_i$ is the consistent functional estimator of $f_{i,0}$.

iii) The approximated errors can also be used to approximate the finite sample distributional properties of $\hat{C}$ by bootstrap.

2.5 Extensions

The results of the Sections above can be used to derive consistent semi-parametric estimators in models of the type :

$$Y_t = a(X_t; \theta) + \Sigma^{1/2} C \varepsilon_t. \tag{2.16}$$

where : $E(Y_t|X_t) = a(X_t; \theta), V(Y_t|X_t) = \Sigma$, C is an orthogonal matrix, and (ε_t) satisfies Assumption A1.

The parameters θ, Σ can be estimated by nonlinear least squares : $\hat{\theta}_T$ is the solution of :

$$\hat{\theta}_T = \arg\min_{\theta} \sum_{t=1}^{T} ||Y_t - a(X_t; \theta)||^2.$$

Then a consistent estimator of Σ is :

$$\hat{\Sigma}_T = \frac{1}{T}[Y_t - a(X_t; \hat{\theta}_T)][Y_t - a(X_t; \hat{\theta}_T)]'.$$

These first-step estimators are used to compute standardized OLS residuals :

$$\hat{u}_t = \hat{\Sigma}_T^{-1/2}[Y_t - a(X_t; \hat{\theta}_T)].$$

The orthogonal matrix C is finally estimated by applying a PML approach on the series of residuals $\hat{u}_t$.

This consistent estimation approach can be applied to either static or dynamic models. In particular it can be used to identify independent shocks in a structural vector autoregressive (SVAR) model [see e.g. Chen, Choi, Escanciano (2012), Moneta et al. (2013). Gourieroux, Monfort (2014)]. In this case the model of interest is :

$$Y_t = \Phi Y_{t-1} + \Sigma^{1/2} C \varepsilon_t.$$

3 Links with the PML literature on ICA

In Section 2, we have derived the asymptotic properties of a PML approach, namely its consistency and its asymptotic normality using the constraints of orthogonal matrix C to solve the identification issue.

There exist other PML estimation methods proposed in the literature on ICA and blind separation of sources. They differ by the identification restrictions which are used,by the possible introduction of auxiliary parameters in the pseudo-log likelihood, by the global or recursive nature of the optimization problem, and by the possible prewhitening of the observed data. Since this literature mainly deals with signal processing and data analysis, there is a focus on the numerical convergence and computational complexity of the algorithm used to optimize the pseudo log-likelihood function. A few papers derive the asymptotic distribution of PML or recursive PML estimators [see e.g. Pham, Garat (1997), Hyvarinen (1997), Ilmonen et al. (2012)], but give no proof of the statistical consistency of the PML estimators. This explains why among the PML methods proposed in the literature and in the softwares several are not statistically consistent. This might also explain practical suggestions such as "In real world problems, it is useful to apply several ICA algorithms, because they may reveal different IC's from the data" [Hyvarinen, Karhunen, Oja (2001), p286]. The aim of this section is to review other PML approaches, to discuss their consistency (or their lack of consistency), and to derive their asymptotic distributional properties.

3.1 Identification by row specific constraints

Let us consider the ICA model :

$$Y_t = C_0 \varepsilon_t, \tag{3.1}$$

with the standard assumptions : C_0 is invertible, the variables $\varepsilon_{1,t}, \ldots, \varepsilon_{n,t}$ are independent, zero mean, and the $\varepsilon_{i,t}, t = 1, \ldots, T$ have a same distribution $f_{i,0}(\varepsilon_i)$. But we do not impose $V(\varepsilon_{i,t}) = 1$. Whereas in Section 2, we have solved the identification issue by assuming an orthogonal matrix C, that is by imposing specific and cross restrictions on the columns $c_i, i = 1, \ldots, n$, of

C, namely $c_i'c_i = 1$ and $c_i'c_j = 0$, the identification issue might also be solved by imposing restrictions only on the rows of C^{-1}, i.e. the SIR2* restrictions :

$$c^i c^{i'} = 1.$$

.

The PML optimization problem becomes :

$$\hat{B}_T = \arg\max_B \sum_{t=1}^{T} \sum_{i=1}^{n} \log g_i(b_i' Y_t), \qquad (3.2)$$
$$s.t.\ b_i' b_i = 1, i = 1, \ldots, n,$$

where B denotes the generic matrix parameter, whose rows are $b_i', i = 1, \ldots, n$.

This problem is numerically very simple, since it is equivalent to n optimization problems, which can be solved independently :

$$\hat{b}_{i,T} = \arg\max_{b_i} \sum_{t=1}^{T} \log g_i(b_i' Y_t) \qquad (3.3)$$
$$s.t.\ b_i' b_i = 1.$$

Such optimization problems are called "one unit algorithms" in the ICA literature [see e.g. Hyvarinen, Oja (2000), Hyvarinen, Karhunen, Oja (2001), Section 8.3].

Proposition 5 : The one unit algorithms provide statistically consistent estimators neither of C_0^{-1}, nor of C_0.

Proof : When the true matrix C_0 is not necessarily orthogonal, the expected interpretation of the pseudo-parameter $B_0 = \lim_{T\to\infty} \hat{B}_T$ is to be equal to C_0^{-1}. Thus we focus below on this expected interpretation.

i) Let us consider the asymptotic first-order conditions corresponding to the optimization problem (3.2). They are :

$$\begin{cases} E_0\left[Y_t \dfrac{d \log g_i}{d\varepsilon}(b_i' Y_t)\right] - \lambda_{i,i} b_i = 0, i = 1, \ldots, n, \\ \\ b_i' b_i = 1, i = 1, \ldots, n \end{cases}$$

We can eliminate the Lagrange multipliers and deduce the equations satisfied by the pseudo-true values only :

$$E_0[Y_t \frac{d \log g_i}{d\varepsilon}(b_i' Y_t)] - E_0[b_i' Y_t \frac{d \log g_i}{d\varepsilon}(b_i' Y_t)] b_i = 0, i = 1, \ldots, n. \qquad (3.4)$$

Let us now check if the solutions in b_i of this system can be the transposed of the rows $c_0^i, i = 1, \ldots, n$ of matrix C_0^{-1}. System (3.4) becomes :

$$E_0[C_0 \varepsilon_t \frac{d \log g_i}{d\varepsilon}(c_0^i Y_t)] - E_0[c_0^i Y_t \frac{d \log g_i}{d\varepsilon}(c_0^{i'} Y_t)] c_0^{i'} = 0, i = 1, \ldots, n$$

$$\Leftrightarrow \quad E_0[C_0 \varepsilon_t \frac{d \log g_i(\varepsilon_{i,t})}{d\varepsilon}] - E_0[\varepsilon_{i,t} \frac{d \log g_i(\varepsilon_{i,t})}{d\varepsilon}] c_0^{i'} = 0, i = 1, \ldots, n$$

$$\Leftrightarrow \quad E_0[\varepsilon_{i,t} \frac{d \log g_i(\varepsilon_{i,t}}{d\varepsilon})](c_{i,0} - c_0^{i'}) = 0, i = 1, \ldots, n,$$

by using the independence between the components of error ε_t and the fact that these components are zero-mean. We deduce that a necessary condition for this one unit algorithm to provide a consistent estimator of C_0^{-1} is :

$$c_{i,0} = c_0^{i'}, i = 1, \ldots, n,$$

that is the orthogonality of the true matrix C_0.

ii) Let us now assume that the true matrix C_0 is orthogonal. We know from the discussion above that the asymptotic FOC are satisfied by $b_i = c_{i,0}$. However, it is also seen that this PML estimator is not statistically consistent in general. Indeed let us choose, as it is standard in the literature, the same pseudo p.d.f. for all indexes i. Then the different optimization problems indexed by i have the same solution, that is $\hat{b}_{i,T} = \hat{b}_T$, independent on i. If they are consistent, their limits are the same c_0, say, and the pseudo-true value of matrix B is $(c_0, c_0, \ldots, c_0)$, which is a noninvertible matrix, that cannot be equal to C_0^{-1}.[7]

[7] In this respect the derivation of the asymptotic distribution of possibly inconsistent one-unit PML estimates have to be considered with care [see e.g. Hyvarinen (1997), Th. 2].

Of course different independent components can be estimated if we change the pseudo pdf in several optimizations of the objective function and run the algorithm using different starting points. Such an approach is rather ad-hoc and does not ensure to find the total number of linearly independent components, contrary to multi-unit methods such as the PML approach developed in Section 2.

3.2 One unit algorithm with row specific constraints and introduction of auxiliary parameters

For the same reason, there is a lack of consistency for more sophisticated one-unit PML approaches [8]. Let us consider the most favorable case of an orthogonal C_0 matrix. We denote by $\sigma_{i,0}^2$ the variance of $\varepsilon_{i,t}$ and by D_0 the diagonal matrix whose diagonal terms are the $\sigma_{i,0}^2$. The variance-covariance matrix of Y_t is $C_0 D_0 C_0'$, which is not constrained. It has been suggested to consider jointly the estimation of matrix C_0 with row specific restrictions on C^{-1}, or equivalently with column restrictions on C, and of the variances $\sigma_{i,0}^2, i = 1, \ldots, n$. The PML estimator is defined on $C^{-1} = C'$ by :

$$(\hat{C}_T, vec(\hat{\sigma}_T^2)) = \arg \max_{C, vec(\sigma^2)} \sum_{t=1}^{T} \sum_{i=1}^{n} [\log g_i(\frac{c_i' Y_t}{\sigma_i}) - \frac{1}{2} \log \sigma_i^2], \tag{3.5}$$

$$s.t.\ c_i' c_i = 1, i = 1, \ldots, n,$$

which is equivalent to n optimizations of smaller dimension considered independently :

$$(\hat{c}_{i,T}, \hat{\sigma}_{i,T}^2) = \arg \max_{c_i, \sigma_i^2} \sum_{t=1}^{T} [\log g_i(\frac{c_i' Y_t}{\sigma_i}) - \frac{1}{2} \log \sigma_i^2], \tag{3.6}$$

$$s.t.\ c_i' c_i = 1.$$

[8] or of more sophisticated versions of such one-unit PML approaches such as the deflationary or the symmetric orthogonalization approaches [see e.g. Hyvarinen, Oja (2000), Hyvarinen, Karhunen, Oja (2001), Sections 8.4.2, 8.4.3]. Indeed these algorithmic methods do not ensure to find all independent components.

It is checked in Appendix A.3.1, that the asymptotic FOC are satisfied by values $c_{i,0}, \sigma_{i,0}^{2*}$, where $c_{i,0}$ is the true value of the i^{th} column of C and $\sigma_{i,0}^{2*}$ differs from the true value $\sigma_{i,0}^2$. Thus, from the FOC, we might expect $\hat{c}_{i,T}$ to be consistent of $c_{i,0}$. However, this approach is not consistent in general for the same reason as in the second part of the proof of Proposition 5. Thus we have the following result :

Proposition 6 : The one-unit algorithm with auxiliary volatility parameter does not provide a statistically consistent estimator of C_0, even if C_0 is an orthogonal matrix.

3.3 Jacobian adjusted PML with row specific constraints and auxiliary parameters

The pseudo-likelihood used in optimization (3.3) is mispecified since the pseudo p.d.f. does not correspond to the true p.d.f., but also since we have not taken into account the Jacobian effect. The Jacobian adjusted PML is the solution of :

$$(\hat{B}_T, vec\hat{\sigma}_T^2) = \arg \max_{C, vec(\sigma^2)} \sum_{t=1}^{T} \sum_{i=1}^{n} \left[\log g_i(\frac{b_i' Y_t}{\sigma_i}) - \frac{1}{2} \log \sigma_i^2 + \log |\det B| \right],$$
$$s.t. b_i' b_i = 1, i = 1, \ldots, n, \tag{3.7}$$

where e_i denotes the i^{th} column of the identity matrix.

This form of objective function has been considered in Pham, Garat (1997), Section 2.A, but without taking into account explicitly the constraints $b_i' b_i = 1, \ldots, n$ in the FOC.[9] Moreover, contrary to the title of their Section 2 : "The ML approach for white sources", they do not really study the properties of the associated PML estimator, but modify the FOC to get covariance restrictions [see their equation (2.1), and our next Section 4 for the analysis of Covariance estimators]. As in subsections 3.1-3.2, the estimator $\hat{B}_T$ solution of the optimization problem (3.7) is not a consistent estimator of C_0^{-1} (see appendix A.3.2), except if C_0 is orthogonal.

[9] even if these restrictions are mentioned p1713 : $\hat{C}_T$ is defined "up to a scaling factor for each of its column".

3.4 Recursive PML approach (under SIR 3)

We have seen in Sections 3.1-3.3 that the one unit identification restrictions SIR2 or SIR2* are not sufficient to get the consistency of the PML estimator of C_0 (or C_0^{-1}), even if the pseudo-likelihood is Jacobian adjusted. Let us now come back to the set of identification restrictions SIR3.

i) **The recursive scheme**

The identification constraints of orthogonality of C can also be introduced in a recursive optimization scheme. Let us consider the same assumptions as in Section 2. In particular C_0 is orthogonal. We can apply a recursive PML approach, called deflation based Fast ICA in the literature [see e.g. Ollila (2010), Reyhani et al. (2012), Ilmonen et al. (2012), Miettinen et al. (2014)]. The recursive PML estimator is derived by a succession of simplified optimization problems.

More precisely at step i, the recursive PML estimators $\hat{c}_1, \ldots, \hat{c}_{i-1}$ have already been derived and the recursive PML estimator $\hat{c}_i$ of c_i is defined as the solution of :

$$\hat{c}_i = \arg\max_{c_i} \sum_{t=1}^{T} \log g_i(c_i' Y_t) \tag{3.8}$$
$$s.t. : c_i' c_i = 1, c_i' \hat{c}_j = 0, j = 1, \ldots, i-1,$$

for $i = 2, \ldots, n$. For $i = 1$, the only constraint is $c_1' c_i = 1$.

ii) **The Gaussian case**

This recursive PML approach has been initially proposed by analogy with principal component analysis (PCA) [see e.g. Lawley, Maxwell (1971), Anderson (1984) for PCA]. PCA is based on a PML approach with Gaussian pseudo-distributions. Taking the standard Gaussian densities for all the densities g_i in formula (2.3), the optimization problem of Section 2 becomes :

$$\max_C \quad -\sum_{t=1}^{T} \sum_{i=1}^{n} (c_i' Y_t)^2$$
$$\text{s.t. } C'C = Id.$$

The objective function can also be written as :

$$
\begin{aligned}
-\sum_{t=1}^{T}\sum_{i=1}^{n} c_i' Y_t Y_t' c_i &= -\sum_{i=1}^{n}[c_i' \sum_{t=1}^{T} Y_t Y_t' c_i] \\
&= -Tr[C' \sum_{t=1}^{T} Y_t Y_t' C] \\
&= -Tr[\sum_{t=1}^{T} Y_t Y_t' CC'] \text{(by commuting within the Trace operator)} \\
&= -Tr(\sum_{t=1}^{T} Y_t Y_t') \text{ (since } CC' = Id).
\end{aligned}
$$

Thus the objective function takes the same value for all orthogonal matrices C. This is the well-known identification problem of matrix C in the Gaussian framework (see the introduction). Then the recursive Gaussian PML is used in PCA to find an easily interpretable matrix C. Indeed the solution of the recursive PML approach is the sequence of unit norm eigenvectors of $\sum_{t=1}^{T} Y_t Y_t'$ associated with the eigenvalues ranked in decreasing order (assuming that there is no multiple eigenvalue).

iii) **Recursive vs global optimization PML estimators**

It can be seen that when the pseudo p.d.f.'s are not Gaussian, the PML estimator of Section 2 and the recursive PML estimator are not necessarily equal in finite sample. For instance let us consider $n = 2$ and parametrize matrix C as[10] :

$$C = \begin{pmatrix} \cos\theta & -\sin\theta \\ \\ \sin\theta & \cos\theta \end{pmatrix}.$$

The PML estimator of θ is the solution of $\max_\theta \sum_{t=1}^{T} \{\log g_1(y_{1,t}\cos\theta +$

[10]This parametrization is valid for a matrix C such that $\det C = 1$.

$y_{2,t}\sin\theta)+\log g_2(-y_{1,t}\sin\theta+y_{2,t}\cos\theta)\}$, whereas the recursive PML estimator of θ is the solution $\max_\theta \sum_{t=1}^{T}[\log g_1(y_{1,t}\cos\theta+y_{2,t}\sin\theta)]$. It is easily seen that the solutions of these optimization problems differ in finite sample (even up to a change of sign on the columns of C). They also have different asymptotic properties. Indeed the conditions of local concavity differ (see assumption $\tilde{A}4$ below). They are respectively :

$$E_0[\frac{d^2\log g_1(\varepsilon_1)}{d\varepsilon^2}+\frac{d^2\log g_2(\varepsilon_2)}{d\varepsilon^2}-\varepsilon_1\frac{d\log g_1(\varepsilon_1)}{d\varepsilon}-\varepsilon_2\frac{d\log g_2(\varepsilon_2)}{d\varepsilon}]<0,$$

$$\text{and}\quad E_0[\frac{d^2\log g_1(\varepsilon_1)}{d\varepsilon^2}-\varepsilon_1\frac{d\log g_1(\varepsilon_1}{d\varepsilon})]<0.$$

The previous identification assumptions A3-A4 are replaced by (see Appendices A.3.3 and A.2.2 for the justification) :

Assumption $\tilde{A}$3 : For any $i=1,\ldots,n-1$, the system :

$$E_0\{\frac{d\log g_i}{d\varepsilon}(c_i'Y_t)[\sum_{j=i}^{n}c_{j,0}\varepsilon_{j,t}-c_i'Y_tc_i-\Sigma_{j<i}\varepsilon_{j,t}c_{j,0}]\}=0,$$

$$c_i'c_i=1, c_i'c_{j,0}=0, j<i, i,j=1,\ldots,n,$$

has the (essentially) unique solution $c_{i,0}$

Assumption $\tilde{A}$4 : local concavity.

The asymptotic objective function is locally concave in a neighbourhood of C_0, if and only if,

$$E_0[\frac{d^2\log g_i(\varepsilon_{i,t})}{d\varepsilon^2}-\varepsilon_{i,t}\frac{d\log g_i(\varepsilon_{i,t})}{d\varepsilon}]<0, i=1,\ldots,n-1.$$

iv) Behaviour of the recursive PML estimator

We prove in Appendix.3.3 that the asymptotic FOC are satisfied by the true values. We deduce the following result :

Proposition 7 : Let us assume that the true matrix C_0 is orthogonal.

i) The PML and recursive PML estimators of C_0 under SIR3 generally differ in finite sample.

ii) Under Assumptions $\tilde{A}3 - \tilde{A}4$ the recursive PML estimator of C_0 is consistent

Additional asymptotic distributional properties of recursive PML estimators have been derived in Ilmonen et al. (2012), Theorem 2.2. and Miettinen et al (2014). In particular it has been realized that Assumption $\tilde{A}3$ is often not satisfied when the same nonlinearities $\frac{d \log g_i}{d\varepsilon}$, independent of i, are introduced in the different steps [see Miettinen et al. (2014), p2].

The finite sample FOC satisfied by the recursive PML estimator are derived in Appendix A-4.2.

4 Generalized Covariance-Estimators

Historically, methods of moments based on cumulants have been the first estimation approaches for ICA proposed in the literature [see e.g. Wiggins (1978), Giannakis et al. (1989), Jutten, Herault (1991), Comon, (1994)].

The use of third and fourth order cumulants explain the lack of robustness of these approaches, whereas the choice of a number of moment restrictions equal to the number of parameters to be estimated explains their lack of efficiency. In this section we develop the Generalized Covariance estimators (GCov) and we show how they can be used in more robust and efficient ways. These Generalized Covariance estimators are serious competitors to PML and recursive PML estimators.

Moreover, it is easily checked that GCov consistent estimators can be derived when there are strictly more sources than sensors [11], whereas the PML and recursive PML approaches require the same number of sources and sensors.

4.1 The covariance conditions

The main theorem for the "essential" identification of matrix C, given in the introduction, is also valid under the weaker assumption that the components

[11] See e.g. Eriksson, Koivunen (2004), Th 5, ii) for conditions on the distribution of the sources to ensure the identification of the independent components

$\varepsilon_{1,t}, \ldots, \varepsilon_{n,t}$ are pairwise independent [see Comon (1994), Th 11, i)]. This assumption can be equivalently written as a set of covariance restrictions :[12]

$$Cov_0[a(c_i'Y_t), b(c_j'Y_t)] = 0, \tag{4.1}$$

for any $i \neq j$, and any (square integrable) functions a, b. These restrictions differ from the standard moment conditions appearing in the generalized method of moments introduced in Hansen (1982), Hansen, Singleton (1982).

Indeed :

$$Cov_0[a(c_i'Y_t), b(c_j'Y_t)] = E_0[a(c_t'Y_t)b(c_j'Y_t)] - E_0[a(c_i'Y_t)]E_0[b(c_j'Y_t)],$$

involves both a moment and a product of moments. This product can be neglected for special choices of function a. For instance, if $a = Id$, a subset of covariance restrictions is :

$$E_0[c_i'Y_t \; b(c_j'Y_t)] = 0, \forall i \neq j, \forall b. \tag{4.2}$$

Such restrictions are for instance considered in Pham, Garat (1997), eq. (2.1) in a just identified case.

We will see below how to use covariance restrictions in a way similar to moment restrictions. Let us note that we need a number of restrictions at least equal to the number of parameters. When we develop the method for an orthogonal matrix C, the number of independent parameters is equal to the number of independent parameters of the skew symmetric matrix of the Cayley's representation, that is $n(n-1)/2$ (see Section 2.1). The order condition is described in the table below :

Table 1 : Order restriction

[12] It is known that the pairwise independence of the sources does not imply their mutual independence when $n \geq 3$. Thus it is also possible to introduce independence restrictions which involve more than two sources. Estimation methods based on cross fourth-order cumulants of the sources have been introduced in the literature, such as the Joint Approximate Diagonalization of Eigenmatrices (JADE) or the Fourth-Order Blind Identification (FOBI) approaches [see e.g. Cardoso Souloumiac (1993), Comon, Mourrain (1996), Hyvarinen, Karhunen, Oja (2001), Chapter 11, and Bonhomme, Robin (2009) for a description of the algorithms, for the asymptotic properties of JADE and quasi JADE estimators]. We do not consider this possibility in our analysis and focus on the efficient use of pairwise independence restrictions.

dimension n	2	3	4	5	6
number of parameters $n(n-1)/2$	1	3	6	10	15

However, more restrictions have to be taken into account in an optimal way, if we want to improve the efficiency of the estimator.

We denote below :

$$Cov_0[\varphi(Y_t,\alpha),\psi(Y_t,\alpha)],$$

the vector with components :

$$Cov_0[\varphi_k(Y_t,\alpha),\psi_k(Y_t,\alpha)], k=1,\ldots,K, K \geq n(n-1)/2,$$

where $\varphi_k(Y_t,\alpha) = a_k[c'_{i_k}(\alpha)Y_t], \psi_k(Y_t,\alpha) = b_k[c'_{j_k}(\alpha)Y_t]$, and α is the vector stacking the lower triangular elements of the skew symmetric matrix of the Cayley's representation of matrix C.

4.2 The generalized covariance (GCov) estimator

The definition of the GCov estimator mimicks the definition of a GMM estimator [Hansen (1982), Hansen, Singleton (1982)]. The GCov estimator is defined as follows from a first step consistent estimator $\tilde{\alpha}_T$. The covariance restriction is first approximated by its sample counterpart expanded around the consistent estimator. We get :

$$\begin{aligned} & \sqrt{T}\widehat{Cov}[\varphi(Y,\tilde{\alpha}_T),\psi(Y,\tilde{\alpha}_T)] \\ \simeq\ & (\frac{\partial}{\partial\alpha'}\widehat{Cov}[\varphi(Y_t,\alpha),\psi(Y_t,\alpha)])_{\alpha=\tilde{\alpha}_T}\sqrt{T}(\tilde{\alpha}_T-\alpha_0)+u_T, \qquad (4.3) \end{aligned}$$

where $u_T = \sqrt{T}\widehat{Cov}[\varphi(Y_t,\alpha_0),\psi(Y,\alpha_0)], E(u_T)=0, V(u_T) = V_{as}[\sqrt{T}\widehat{Cov}(\varphi(Y_t,\alpha_0),\psi(Y_t,\alpha_0))] \equiv \Sigma_0$, say.

System (4.3) is asymptotically a linear model with respect to α_0, which can be estimated by Feasible Generalized Least Squares (Feasible GLS). The GCov estimator is defined as :

$$\hat{\alpha}_T = \tilde{\alpha}_T - \{\frac{\partial}{\partial\alpha}\widehat{Cov}'[\varphi(Y,\tilde{\alpha}_T),\psi(Y,\hat{\alpha}_T)]\hat{\Sigma}^{-1}\frac{\partial}{\partial\alpha'}\widehat{Cov}[\varphi(Y,\hat{\alpha}_T),\psi(Y,\tilde{\alpha}_T)]\}^{-1}$$
$$\frac{\partial}{\partial\alpha}\widehat{Cov}'[\varphi(Y,\tilde{\alpha}_T),\psi(Y,\tilde{\alpha}_T)]\hat{\Sigma}^{-1}\widehat{Cov}[\varphi(Y,\tilde{\alpha}_T),\psi(Y,\tilde{\alpha}_T)], \tag{4.4}$$

where $\hat{\Sigma}$ is a consistent estimator of Σ_0.
Proposition 8 : The GCov estimator is consistent, asymptotically normal with :

$$Cov_{as}[\sqrt{T}(\hat{\alpha}_T - \alpha_0)]$$
$$= \{\frac{\partial}{\partial\alpha}Cov_0'[\varphi(Y_t,\alpha_0),\psi(Y_t,\alpha_0)]\Sigma_0^{-1}\frac{\partial}{\partial\alpha'}Cov_0[\varphi(Y_t,\alpha_0),\psi(Y_t,\alpha_0)]\}^{-1}.$$

The proof of these asymptotic properties and of the optimality of this estimator are similar to the ones for GMM [see e.g. Gourieroux, Monfort (1995), Sections 9.5.2, 9.5.3, and Property 9.11]. The matrices involved in the expression of $\hat{\alpha}_T$, or of its asymptotic distribution, can be explicited as follows :

First we have :

$$\begin{aligned}
&\frac{\partial}{\partial\alpha'}Cov_0[\varphi_k(Y_t,\alpha_0),\psi_k(Y_t,\alpha_0)] \\
=\ & Cov_0[\frac{\partial}{\partial\alpha'}\varphi_k(Y_t,\alpha_0),\psi_k(Y_t,\alpha_0)] + Cov_0[\frac{\partial\psi_k}{\partial\alpha'}(Y_t,\alpha_0),\varphi_k(Y_t,\alpha_0)] \\
=\ & \frac{\partial c_{i_k}}{\partial\alpha'}(\alpha_0)Cov_0[Y_t'\frac{da_k(\varepsilon_{i_k,t})}{d\varepsilon},b_k(\varepsilon_{j_k,t})] \\
+\ & \frac{\partial c_{j_k}}{\partial\alpha'}(\alpha_0)Cov_0[Y_t'\frac{db_k(\varepsilon_{j_k,t})}{d\varepsilon},a_k(\varepsilon_{i_k,t})] \\
=\ & \frac{\partial c_{i_k}(\alpha_0)}{\partial\alpha'}c_{j_k}'(\alpha_0)Cov_0[\varepsilon_{j_k,t}\frac{da_k(\varepsilon_{i_k,t})}{d\varepsilon},b_k(\varepsilon_{j_k,t})] \\
+\ & \frac{\partial c_{j_k}(\alpha_0)}{\partial\alpha'}c_{ik}'Cov_0(\varepsilon_{i_k,t}\frac{db_k(\varepsilon_{j_k,t})}{d\varepsilon},a_k(\varepsilon_{i_k,t})],
\end{aligned} \tag{4.5}$$

by applying the independence assumption.

Second, the elements of matrix Σ_0 can be explicited (see Appendix 6). We get :

$$\begin{aligned}\sigma_{k,l} &= Cov_{as}\{\sqrt{T}\widehat{Cov}[\varphi_k(Y,\alpha_0),\psi_k(Y,\alpha_0)],\sqrt{T}\widehat{Cov}[\varphi_l(Y,\alpha_0),\psi_l(Y,\alpha_0)]\} \\ &= Cov_0\{[\varphi_k(Y_t,\alpha_0)-E_0\varphi_k(Y_t,\alpha_0)][\psi_k(Y_t,\alpha_0)-E_0\psi_k(Y_t,\alpha_0)], \\ &\quad [\varphi_l(Y_t,\alpha_0)-E_0\varphi_l(Y_t,\alpha_0)][\psi_l(Y_t,\alpha_0)-E_0\psi_l(Y_t,\alpha_0)]\}. \qquad (4.6)\end{aligned}$$

They are consistently estimated by replacing the theoretical covariances by their sample counterparts, the parameters by their estimates and the errors by the associated residuals.

Remark : Another type of covariance (or correlation) based estimator has been proposed in Bach, Jordan (2002). Loosely speaking, they propose to estimate α by minimizing

$$\rho^*(\alpha) = \max_{k=1,\ldots,K} |\hat{corr}(\varphi_k(Y,\alpha),\psi_k(Y,\alpha)|,$$

by analogy with the standard canonical correlation analysis. They do not provide the asymptotic distributional properties of this estimator, but it is known that it is less efficient asymptotically than the generalized covariance estimator introduced above.

5 Concluding Remarks

There is a huge literature proposing semi-parametric estimation methods for the mixing matrix in models with independent component analysis. These methods are essentially pseudo maximum likelihood approaches, or methods based on covariance restrictions. However the standard literature focuses on the numerical properties of these methods such as their numerical convergence, but generally neglects their statistical properties : statistical convergence and asymptotic distribution. The aim of our paper was to consider these statistical properties. In particular

i) We show that the one unit PML approaches, often used in practice, are not statistically consistent.

ii) We derive the necessary and sufficient identification conditions for multi-unit PML and recursive PML approaches, whereas only sufficient conditions have been derived in the literature.

iii) We show that the multi-unit PML approaches under the constraint of orthogonal mixing matrix are consistent and we provide the asymptotic distribution of the multi-unit PML estimator.

iv) We explain how to improve the efficiency of covariance based estimators and analyze the properties of the generalized covariance estimators.

v) We discuss the links between the empirical likelihood estimators and the covariance estimators (in Appendix 7).

The PML and covariance based approaches are largely used in practice even if they do not allow to reach the (semi-) parametric efficiency bound. Semi-parametric efficient methods have been introduced in the more theoretical literature (see the review in Appendix 8).

All these methods are more difficult to implement than the PML and covariance based approaches. There is clearly a trade-off between statistical efficiency and numerical simplicity [see the comparison of performances in Figure 1 of Chen, Bickel (2005)]. They can be less robust, requiring for instance mutually independent errors, whereas the generalized covariance estimators are consistent under pairwise independence only.

Moreover, they are often difficult to extend to a dynamic framework, especially to the consistent estimation of the moving average parameters $C_j, j = -\infty, \ldots, +\infty$, from observations of a stationary process satisfying :

$$Y_t = \sum_{j=-\infty}^{\infty} C_j \varepsilon_{t-j}$$

[see e.g. Gourieroux, Monfort (2014), Gourieroux, Jasiak (2015), for the estimation of such parameters by covariance estimators].

R E F E R E N C E S

Amari, S., and J., Cardoso (1997) : "Blind Source Separation. Semi-Parametric Structural Approach", IEEE Trans. on Signal Processing, 45, 2692-2700.

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

Bach, F., and M., Jordan (2002) : "Kernel Independent Component Analysis", J. Machine Learning Res., 3, 1-48.

Bell, A., and T., Sejnowski (1995) : "An Information-Maximization Approach to Blind Separation and Blind Deconvolution", Neural Computation, 7, 1129-1159.

Belouchrani, A., Abed-Meraim, K., Cardoso, J.F., and E., Moulines (1997) : "A Blind Source Separation Technique Using Second-Order Statistics", IEEE Trans. On Signal Processing, 45, 434-444.

Bonhomme, S., and J.M., Robin (2009) : "Consistent Noisy Independent Component Analysis", Journal of Econometrics, 149, 12-25.

Cardoso, J. (1989) : "Source Separation Using Higher Moments", in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2109-2112.

Cardoso, J. (1999) : "High-order Contrasts for Independent Component Analysis", Neural Comput., 11, 157-192.

Cardoso, J., and B., Hvam Laheld (1996) : "Equivariant Adaptive Source Separation", IEEE Tran. on Signal Processing, 44, 3017-3030.

Cardoso, J., and A., Souloumiac (1993) : "Blind Beamforming for Non Gaussian Signals", IEE. Proceedings, F, 140, 362-370.

Chen, A., and P., Bickel (2005) : "Consistent Independent Component Analysis and Prewhitening", IEEE Transactions on Signal Processing, 53,

3625-3632.

Chen, A., and Bickel, P. (2006) : "Efficient Independent Component Analysis", The Annals of Statistics, 34, 2825-2855.

Chen, B., Choi, J., and J.C., Escanciano (2012) : "Testing for Fundamental Moving Average Representation", DP Indiana University.

Cichocki, A., and S., Amari (2006) : "Adaptive Blind Signal and Image Processing", Wiley, Chichester.

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

Comon, P. and B., Mourrain (1996) : "Decomposition of Quantics in Sums of Powers of Linear Forms", Signal Processing, 53, 93-107.

Comon, P., Jutten, C., and J., Herault (1991) : "Blind Separation of Sources, Part II : Problems Statement", Signal Processing, 24, 11-20.

Degerine, S., and R., Malki (2000) : "Second-Order Blind Separation of Sources Based on Canonical Partial Innovations", IEEE Trans. on Signal Processing, 48, 629-641.

Eloyan, A., and S., Ghosh (2011) : "Smooth Density Estimation with Moment Constraints Using Mixture Distributions", Journal of Nonparametric Statistics, 23, 513-531.

Eriksson, J., and V., Koivunen (2004) : "Identifiability, Separability and Uniqueness of Linear ICA Models", IEEE Signal Processing Letters, 11, 601-604.

Gourieroux, C., and J., Jasiak (2015) : "Semi-Parametric Estimation of Noncausal Vector Autoregression", CREST DP.

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

Gourieroux, C., and A., Monfort (2014) : "Revisiting Identification and Estimation in Structural VARMA Models", CREST DP.

Hansen, L. (1982) : "Large Sample Properties of Generalized Method of Moment Estimators", Econometrica, 50, 1029-1054.

Hansen, L., and K., Singleton (1982) : "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models", Econometrica, 50, 1269-1286.

Hastie, T., and R., Tibshirani (2002) : "Independent Component Analysis Through Product Density Estimators", DP Stanford University.

Hyvarinen, A. (1997) : "Independent Component Analysis by Minimization of Mutual Information", Helsinki University of Technology.

Hyvarinen, A. (1999) : "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis", IEEE Transactions on Neural Networks, 10, 626-634.

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

Hyvarinen, A., and E., Oja (1997) : "A Fast Fixed Point Algorithm for Independent Component Analysis", Neural Computation, 9, 1483-1492.

Hyvarinen, A., and E., Oja (2000) : "Independent Component Analysis : Algorithms and Applications", Neural Networks, 13, 411-430.

Ilmonen, P. (2013) : "On Asymptotic Properties of the Scatter Matrix Based Estimates for Complex Valued Independent Component Analysis", Probability Letters, 83, 1219-1226.

Ilmonen, P., Nordhausen, K., Oja, H., and E., Ollila (2012) : "On Asymptotics of ICA Estimators and Their Performance Indices", DP.

Ilmonen, P., and D., Paindaveine (2015) : "Semi-Parametrically Efficient Inference Based on Signed Ranks in Symmetric Independent Component Models", The Annals of Statistics, 39, 2448-2476.

Jennrich, R (1969) : "Asymptotic Properties of Nonlinear Least Squares Estimators", The Annals of Mathematical Statistics, 40, 633-643.

Jutten, C., and J., Herault (1991) : "Blind Separation of Sources. Part 1 : An Adaptive Algorithm Based on Neuromimetic Structure", Signal Processing, 24, 1-10.

Kagan, A., Linnick, Y., and C., Rao (1973) : "Characterization Problems in Mathematical Statistics", Ser. Probability and Mathematical Statistics, New-York, Wiley.

Kaiser, H. (1958) : "The Varimax Criterion for Analytic Rotation in Factor Analysis", Psychometrika, 23, 187-200.

Kitamura, Y., Tripathi, G., and H., Ahn (2004) : "Empirical Likelihood-Based Inference in Conditional Moment Restriction Models", Econometrica, 72, 1667-1714.

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

Miettinen, J., Nordhausen, K., Oja, H., and S., Taskinen (2014) : "Deflation-Based Fast ICA with Adaptive Choices of Nonlinearities", IEEE Transactions on Signal Processing, 1-9.

Moneta, A., Entner, D., Hoyer, P., and A., Coad (2013) : "Causal Inference by Independent Component Analysis: Theory and Applications", Oxford Bulletin of Economics and Statistics, 75, 705-730.

Oja, M., Sirkia, S., and J., Eriksson (2006) : "Scatter Matrices and Independent Component Analysis", Austrian Journal of Statistics, 35, 175-189.

Ollila, E. (2010) : "The Deflation-Based FastICA Estimator : Statistical Analysis Revisited", IEEE Transaction in Signal Processing, 58, 175-189.

Pham, D., and P., Garat (1997) : "Blind Separation of Mixture of Independent Sources Through a Quasi-Maximum Likelihood Approach", IEEE Transactions on Signal Processing, 45, 1712-1725.

Reyhani, N., Ylipaavalniemi, J., Vigario, R., and O., Erkki (2012) : "Consistency and Asymptotic Normality of FastICA and Bootstrap FastICA", Signal Processing, 92, 1767-1778.

Samarov, A., and A., Tsybakov (2004) : "Nonparametric Independent Component Analysis", Bernoulli, 10, 565-582.

Tong, L., Soon, V., Huang, Y., and R., Liu (1990) : "Amuse : A New Blind Identification Algorithm", in Proc. IEEE ISCAS, 1784-1787, New-Orleans, May.

Tong, L., Soon, V., Huang, Y., and R., Liu (1991) : "Indeterminacy and Identifiability of Blind Identification", IEEE Trans. Signal Processing, 38, 499-509.

Vlassis, N., and Y., Motomura (2001) : "Efficient Source Adaptivity in Independent Component Analysis", IEEE Trans. Neural Networks, 12, 559-565.

Wei, T. (2014) : "The Convergence and Asymptotic Analysis of the Generalized Symmetric Fast ICA Algorithm", DP University of Lille.

Wiggins, R. (1978) : "Minimum Entropy Deconvolution", Geoexploration, 16, 21-35.

Appendix 1

Expansion of the Cayley's Representation

of an Orthogonal Matrix

Let us perform the second-order expansion of $C(A) = (Id+A)(Id-A)^{-1}$ with respect to A. Let us denote $A = A_0 + \Delta A$, where ΔA is a small skew-symmetric matrix. We have :

$$
\begin{aligned}
& C(A) \\
= \; & (Id + A_0 + \Delta A)(Id - A_0 - \Delta A)^{-1} \\
= \; & (Id + A_0 + \Delta A)\{[Id - \Delta A(Id - A_0)^{-1}](Id - A_0)\}^{-1} \\
= \; & (Id + A_0 + \Delta A)(Id - A_0)^{-1}[Id - \Delta A(Id - A_0)^{-1}]^{-1} \\
= \; & [C(A_0) + \Delta A(Id - A_0)^{-1}][Id + \Delta A(Id - A_0)^{-1} + \Delta A(Id - A_0)^{-1}\Delta A(Id - A_0)^{-1}] + o||\Delta A||^2 \\
= \; & C(A_0) + \{\Delta A(Id - A_0)^{-1} + C(A_0)\Delta A(Id - A_0)^{-1}\} \\
+ \; & \Delta A(Id - A_0)^{-1}\Delta A(Id - A_0)^{-1} + C(A_0)\Delta A(Id - A_0)^{-1}\Delta A(Id - A_0)^{-1} + o||\Delta A||^2 \\
= \; & C(A_0) + [Id + C(A_0)]\Delta A(Id - A_0)^{-1} \\
+ \; & [Id + C(A_0)]\Delta A(Id - A_0)^{-1}\Delta A(Id - A_0)^{-1} + o(||\Delta A||)^2.
\end{aligned}
$$

Since :

$$
\begin{aligned}
& Id + C(A_o) \\
= \; & Id + (Id + A_0)(Id - A_0)^{-1} \\
= \; & (Id - A_0 + Id + A_0)(Id - A_0)^{-1} \\
= \; & 2(Id - A_0)^{-1},
\end{aligned}
$$

we get also :

$$C(A) = C(A_0) + 2(Id - A_0)^{-1}\Delta A(Id - A_0)^{-1}$$

$$+2(Id - A_0)^{-1}\Delta A(Id - A_0)^{-1}\Delta A(Id - A_0)^{-1} + o(||\Delta A||)^2.$$

We deduce the expansion of the transpose $C'(A)$ by using the equalities : $\Delta A' = -\Delta A, A_0' = -A_0$:

$$C'(A) = C'(A_0) - 2(Id + A_0)^{-1}\Delta A(Id + A_0)^{-1}$$

$$-2(Id + A_0)^{-1}\Delta A(Id + A_0)^{-1}\Delta A(Id + A_0)^{-1} + o(||\Delta A||)^2,$$

with $C'(A) = (Id + A)^{-1}(Id - A)$.

We also deduce :

$$\begin{aligned}
C'(A_0)Y_t &= C'(A_0)C(A_0)\varepsilon_t \\
&= \varepsilon_t - 2(Id + A_0)^{-1}\Delta A(Id + A_0)^{-1}C(A_0)\varepsilon_t \\
&- 2(Id + A_0)^{-1}\Delta A(Id + A_0)^{-1}\Delta A(Id + A_0)^{-1}C(A_0)\varepsilon_t + o(||\Delta A||)^2 \\
&= \varepsilon_t - 2(Id + A_0)^{-1}\Delta A(Id - A_0)^{-1}\varepsilon_t \\
&- 2(Id + A_0)^{-1}\Delta A(Id + A_0)^{-1}\Delta A(Id - A_0)^{-1}\varepsilon_t + o(||\Delta A||)^2.
\end{aligned}$$

Appendix 2

Local Concavity of the Pseudo Log-Likelihood Function

A.2.1 PML estimator (with SIR3)

i) Let us first explicit the second-order expansion of the asymptotic objective function without taking into account the constraints of orthogonal C matrix. We introduce the notation $c_i = c_{i,0} + \delta_i$, where δ_i is small. We get :

$$\begin{aligned} L_\infty(\delta) &= E_0[\sum_{i=1}^{n} \log g_i(c_i' Y_t)] \\ &\simeq E_0\{\sum_{i=1}^{n}[\log g_i(c_{i,0}' Y_t) + \frac{d \log g_i}{d\varepsilon}(c_{i,0}' Y_t)\delta_i' Y_t + \frac{1}{2}\frac{d^2 \log g_i}{d\varepsilon^2}(c_{i,0}' Y_t)(\delta_i' Y_t)^2\}. \end{aligned}$$

Since $Y_t = \sum_{j=1}^{n} c_{j,0}\varepsilon_{j,t}$, we deduce :

$$\begin{aligned} L_\infty(\delta) &\simeq E_0[\sum_{i=1}^{n} \log g_i(\varepsilon_{i,t})] + \sum_{i=1}^{n}\sum_{j=1}^{n} E_0\left[\frac{d \log g_i(\varepsilon_{i,t})}{d\varepsilon}\varepsilon_{j,t}\right]\delta_i' c_{j,0} \\ &+ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\{E_0[\frac{d^2 \log g_i(\varepsilon_{i,t})}{d\varepsilon^2}\varepsilon_{j,t}\varepsilon_{k,t}]\delta_i' c_{j,0}\delta_i' c_{k,0} \\ &= E_0[\sum_{i=1}^{n} \log g_i(\varepsilon_{i,t})] + \sum_{i=1}^{n} E_0[\frac{d \log g_i(\varepsilon_{i,t})}{d\varepsilon}\varepsilon_{i,t}]\delta_i' c_{i,0} \\ &+ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} E_0[\frac{d^2 \log g_i(\varepsilon_{i,t})}{d\varepsilon^2}\varepsilon_{j,t}^2](\delta_i' c_{j,0})^2, \end{aligned}$$

by using the independence property.

Since :

$$E_0[\frac{d^2 \log g_i(\varepsilon_{i,t})}{d\varepsilon^2}\varepsilon_{j,t}^2]$$

$$= E_0[\frac{d^2 \log g_i(\varepsilon_{i,t})}{d\varepsilon^2}]E_0(\varepsilon_{j,t}^2)$$

$$= E_0[\frac{d^2 \log g_i(\varepsilon_{i,t})}{d\varepsilon^2}], \text{ if } i \neq j,$$

we get :

$$\begin{aligned} L_\infty(\delta) &\simeq E_0[\sum_{i=1}^n \log g_i(\varepsilon_{i,t})] + \sum_{i=1}^n E_0[\frac{d \log g_i(\varepsilon_{i,t})}{d\varepsilon}\varepsilon_{i,t}]\delta_i' c_{i,0} \\ &+ \frac{1}{2}\sum_{i=1}^n E_0[\frac{d^2 \log g_i(\varepsilon_{i,t})}{d\varepsilon^2}\varepsilon_{it}^2](\delta_i' c_{i,0})^2 \\ &+ \frac{1}{2}\sum_{i=1}^n E_0[\frac{d^2 \log g_i(\varepsilon_{i,t})}{d\varepsilon^2}][\delta_i'\delta_i - (\delta_i' c_{i,0})^2], \end{aligned}$$

with $\sum_{j=1}^n (\delta_i' c_{j,0})^2 = \sum_{j=1}^n (\delta_i' c_{j,0} c_{j,0}' \delta_i) = \delta_i' C_0 C_0' \delta_i = \delta_i'\delta_i$.

This expansion of the objective function involves the n^2 infinitesimal coordinates $\Delta_{i,j} \equiv -c_{i,0}'\delta_j, i, j = 1, \ldots, n$, which are submitted to the $n(n+1)/2$, restrictions of orthogonal C matrix.

ii) Let us now expand the restrictions of orthogonal matrix C .

They are equivalent to :

$$c_{i,0}'\delta_j + c_{j,0}'\delta_i + \delta_i'\delta_j = 0, i \leq j.$$

These equations show that $c_{i,0}'\delta_i = -\frac{1}{2}\delta_i'\delta_i$ and $c_{i,0}'\delta_j + c_{j,0}'\delta_i = -\delta_i'\delta_j$ are of second-order. Let us now eliminate the negligible terms in the expansion of $L_\infty(\delta)$. We get :

$$
\begin{aligned}
L_\infty(\delta) &\simeq E_0[\sum_{i=1}^{n} \log g_i(\varepsilon_{i,t})] - \frac{1}{2}\sum_{i=1}^{n} E_0[\frac{d\log g_i(\varepsilon_{i,t})}{d\varepsilon}\varepsilon_{i,t}]\delta_i'\delta_i \\
&+ \frac{1}{2}\sum_{i=1}^{n} E_0[\frac{d^2\log g_i(\varepsilon_{i,t})}{d\varepsilon^2}\varepsilon_{i,t}^2](\delta_i' c_{i,0})^2 \\
&+ \frac{1}{2}\sum_{i=1}^{n} E_0[\frac{d^2\log g_i(\varepsilon_{i,t})}{d\varepsilon^2}][\delta_i'\delta_i - (\delta_i' c_{i,0})^2] \\
&\simeq E_0[\sum_{i=1}^{n} \log g_i(\varepsilon_{i,t})] + \frac{1}{2}\sum_{i=1}^{n}\sum_{j\neq i}\{E_0[\frac{d^2\log g_i(\varepsilon_{i,t})}{d\varepsilon^2} - \frac{d\log g_i(\varepsilon_{i,t})}{d\varepsilon}\varepsilon_{i,t}](\delta_i' c_{j,0})^2 \\
&\simeq E_0[\sum_{i=1}^{n} \log g_i(\varepsilon_{i,t})] \\
&+ \frac{1}{2}\sum_{i=1}^{n}\sum_{j>i} E_0[\frac{d^2\log g_i(\varepsilon_{i,t})}{d\varepsilon^2} + \frac{d^2\log g_j(\varepsilon_{j,t})}{d\varepsilon^2} - \frac{d\log g_i(\varepsilon_{i,t})}{d\varepsilon}\varepsilon_{i,t} - \frac{d\log g_j(\varepsilon_{j,t}}{d\varepsilon}\varepsilon_{j,t}](\delta_i' c_{j,0})^2.
\end{aligned}
$$

since $\delta_i' c_{j,0} \simeq -\delta_j' c_{i,0}$

This expansion involves the $n(n-1)/2$ functionally independent components of Δ at order 1. Then the condition for local concavity follows.

A. 2.2 Recursive PML estimator (under SIR3)

Let us now consider the conditions for the recursive PML estimator. At iteration i, the expansion of the asymptotic objective function becomes :

$$
\begin{aligned}
L_\infty(\delta_i) &\simeq E_0 \log g_i(\varepsilon_{i,t}) + E_0\left(\frac{d\log g_i(\varepsilon_{i,t})}{d\varepsilon}\varepsilon_{i,t}\right)\delta_i' c_{i,0} \\
&+ \frac{1}{2}\sum_{j=1}^{n} E_0\left[\frac{d^2\log g_i(\varepsilon_{i,t})}{d\varepsilon^2}\varepsilon_{j,t}^2\right](\delta_i' c_{j,0})^2,
\end{aligned}
$$

whereas the restrictions for orthogonal matrix C are equivalent to :

$$\delta_i' c_{j,0} = 0, \forall j < i,\ 2\delta_i' c_{i,0} + \delta_i' \delta_i = 0.$$

Since $\delta_i' c_{i,0} = -(1/2)\delta_i' \delta_i$ is of order 2, the expansion of the objective function becomes :

$$L_\infty(\delta_i) \simeq E_0 \log g_i(\varepsilon_{i,t}) + \frac{1}{2}\{E_0[\frac{d^2 \log g_i(\varepsilon_{i,t})}{d\varepsilon^2}] - E_0[\frac{d \log g_i(\varepsilon_{i,t})}{d\varepsilon}\varepsilon_{i,t}]\}\delta_i' \delta_i.$$

Thus the condition for local concavity is :

$$E_0[\frac{d^2 \log g_i(\varepsilon_{i,t})}{d\varepsilon^2}] - E_0[\frac{d \log g_i(\varepsilon_{i,t})}{d\varepsilon}\varepsilon_{i,t}] < 0,$$

and has to be written for $i = 1, \ldots, n-1$.

Appendix 3

Consistency of the PML with Auxiliary Parameters and Recursive PML Estimator.

A.3.1 PML with auxiliary parameters

The asymptotic FOC for optimization problem (3.5) are :

$$\begin{cases} E_0[\dfrac{\partial \log g_i}{d\varepsilon}(\dfrac{c_i'Y_t}{\sigma_i})\dfrac{Y_t}{\sigma_i}] - \lambda_{i,i}c_i = 0, \\ \\ E_0[\dfrac{\partial \log g_i}{\partial\varepsilon}(\dfrac{c_i'Y_t}{\sigma_i})\dfrac{c_i'Y_t}{\sigma_i^2} + \dfrac{1}{\sigma_i}] = 0. \end{cases}$$

They are equivalent to :

$$\begin{cases} E_0[\dfrac{d \log g_i}{d\varepsilon}(\dfrac{c_i'Y_t}{\sigma_i})\dfrac{c_{j,0}'Y_t}{\sigma_i}] - \lambda_{i,i}c_{j,0}'c_i = 0, \forall j \neq i, \\ \\ E_0[\dfrac{d \log g_i}{d\varepsilon}(\dfrac{c_i'Y_t}{\sigma_i})\dfrac{c_{i,0}'Y_t}{\sigma_i}] - \lambda_{i,i}c_{i,0}'c_i = 0, \\ \\ E_0[\dfrac{d \log g_i}{d\varepsilon}(\dfrac{c_i'Y_t}{\sigma_i})\dfrac{c_i'Y_t}{\sigma_i^2} + \dfrac{1}{\sigma_i}] = 0. \end{cases}$$

The first subsystem is satisfied for $c_i = c_{i,0}$ and any value of σ_i. Then the third subsystem is used to find the appropriate value of σ_i, which is generally different from $\sigma_{i,0}$, whereas the second equation fixes the asymptotic value of the Lagrange multiplier.

A.3.2 Jacobian adjusted PML with auxiliary parameters

The constrained optimization problem is :

$$\begin{cases} \max_B \sum_{t=1}^{T}[\sum_{i=1}^{n}\{\log g_i(\dfrac{b_i'Y_t}{\sigma_i}) + \log\det B - \dfrac{1}{2}\log\sigma_i^2], \\ \\ \text{s.t.: } b_i'b_i = 1, i = 1, \ldots, n, \end{cases}$$

where $B = \begin{pmatrix} b_1' \\ \vdots \\ b_n' \end{pmatrix}$.

The associate asymptotic criterion is :

$$\sum_{i=1}^{n} \{E_0[\log g_i(\frac{b_i' Y_t}{\sigma_i}) + \log \det B - \frac{1}{2} \log \sigma_i^2],$$

and the asymptotic FOC for b_i are :

$$E_0[Y_t \frac{d \log_i}{d\varepsilon}(\frac{b_i' Y_t}{\sigma_i})] - \lambda_{i,i} b_i + b^i = 0, i = 1, \ldots, n,$$

where the derivative of $\log \det B$ with respect to B is $(B^{-1})'$, b^i denotes the i^{th} column of B^{-1}, and $\lambda_{i,i}/2$ the Lagrange multiplier corresponding to the restrictions $b_i' b_i = 1$.

Let us now check if C_0^{-1} is solution of these asymptotic FOC. These FOC become :

$$\sum_{j=1}^{n} c_{0,j} E_0[\varepsilon_{j,t} \frac{d \log g_i(\varepsilon_{i,t}/\sigma_i)}{d\varepsilon}] - \lambda_{i,i} c_0^i + c_{0,i} = 0$$

$$\text{or } c_{0,i} \{E_0[\varepsilon_{i,t} \frac{d \log g_i(\varepsilon_{i,t}/\sigma_i)}{d\varepsilon}] + 1] - \lambda_{i,i} c_0^i = 0.$$

Then, we have to distinguish two cases :

i) If the matrix C_0 is not orthogonal, these FOC are not satisfied.

ii) If the matrix C_0 is orthogonal, we have $c_0^i = c_{0,i}$. The constraints of the optimization problem are satisfied and the FOC above provides the value of the (asymptotic) Lagrange multiplier for a given value of σ_i :

$$\lambda_{i,i} = E_0[\varepsilon_{i,t} \frac{d \log g_i(\varepsilon_{i,t}/\sigma_i)}{d\varepsilon}] + 1.$$

The value of σ_i is deduced from the asymptotic FOC for σ_i :

$$E_0[-\frac{\varepsilon_{i,t}}{\sigma_i^2} \frac{d \log g_i(\varepsilon_{i,t}/\sigma_i)}{d\varepsilon}] - \frac{1}{\sigma_i} = 0.$$

A.3.3 Asymptotic FOC for the recursive PML estimator

Let us denote by $\lambda_{i,i}/2, \lambda_{i,j}, j < i$, the Lagrange multipliers associated with restrictions $c_i'c_i = 1, c_i'c_{j,0} = 0, j < i$. The derivative of the asymptotic Lagrangian associated with the optimization problem (3.8) provides the system :

$$E_0[Y_t \frac{d \log g_i}{d\varepsilon}(c_i'Y_t)] - \lambda_{i,i}c_i - \Sigma_{j<i}\lambda_{i,j}c_{j,0} = 0,$$

with $c_i'c_i = 1, c_i'c_{j,0} = 0, j < i$.

By multiplying the first equation by c_i', and by $c_{j,0}', j = 1, \ldots, i-1$, and using the orthogonality conditions, including $c_{j,0}'c_{k,0} = 0, k \neq j \leq i-1$ $c_{j,0}'c_{j,0} = 1, j = 1, \ldots, i-1$, we get :

$$\lambda_{i,i} = E_0[c_i'Y_t \frac{d \log g_i}{d\varepsilon}(c_i'Y_t)], \lambda_{i,j} = E_0[c_{j,0}'Y_t \frac{d \log g_i}{d\varepsilon}(c_i'Y_t)] = E_0[\varepsilon_{j,t} \frac{d \log g_i}{d\varepsilon}(c_i'Y_t)].$$

Thus the system becomes :

$$E_0\{\frac{d \log g_i}{d\varepsilon}(c_i'Y_t)[\sum_{j=i}^{n} c_{j,0}\varepsilon_{j,t} - c_i'Y_tc_i - \Sigma_{j<i}\varepsilon_{j,t}c_{j,0}]\} = 0,$$

$$c_i'c_i = 1, c_i'c_{j,0} = 0, j < i.$$

We see that the true $c_{i,0}$ is solution of this system. Indeed for $c_i = c_{i,0}$ the first subsystem becomes :

$$E_0\{\frac{d \log g_i(\varepsilon_{i,t})}{d\varepsilon}[\sum_{j=i}^{n} c_{j,0}\varepsilon_{j,t} - \varepsilon_{i,t}c_{i,0}]\}$$

$$= \quad E_0\{\frac{d \log g_i(\varepsilon_{i,t})}{d\varepsilon}[c_{i,0}\varepsilon_{i,t} - c_{i,0}\varepsilon_{i,t}]\} = 0.$$

We deduce from the computation above the identification assumption $\tilde{A}3$.

Appendix 4

Asymptotic Distributions of the PML and Recursive PML Estimators

A.4.1 PLM estimator

Let us consider the finite sample FOC (2.9) :

$$\begin{cases} \sum_{t=1}^{T} \hat{c}_j' Y_t \frac{d \log g_i}{d\varepsilon}(\hat{c}_i' Y_t) - \sum_{t=1}^{T} \hat{c}_i' Y_t \frac{d \log g_j}{d\varepsilon}(\hat{c}_j' Y_t) = 0, i < j, \\ \hat{c}_i' \hat{c}_j = 0, i < j, \hat{c}_i' \hat{c}_i = 1, i = 1, \ldots, n. \end{cases} \tag{a.1}$$

Let us denote $\delta_i = \hat{c}_i - c_{i,0}$ the difference between the PML estimator and the true value. A first-order expansion of the equations in (a.1) gives :

$$\begin{cases} \sum_{t=1}^{T} (c_{j,0}' + \delta_j') Y_t \frac{d \log g_i}{d\varepsilon}(c_{i,0}' Y_t) + \sum_{t=1}^{T} c_{j,0}' Y_t \frac{d^2 \log g_i}{d\varepsilon^2}(c_{i,0}' Y_t) \delta_i' Y_t \\ - \sum_{t=1}^{T} (c_{i,0}' + \delta_i') Y_t \frac{d \log g_j}{d\varepsilon}(c_{j,0}' Y_t) - \sum_{t=1}^{T} c_{i,0}' Y_t \frac{d^2 \log g_j}{d\varepsilon^2}(c_{j,0}' Y_t) \delta_j' Y_t \simeq 0, i < j, \\ c_{i,0}' \delta_j + c_{j,0}' \delta_i \simeq 0, i < j, c_{i,0}' \delta_i \simeq 0, i = 1, \ldots, n. \end{cases}$$

Let us focus on the first subsystem. This subsystem is equivalent to :

$$\begin{aligned} & \sum_{t=1}^{T} [\varepsilon_{j,t} \frac{d \log g_i(\varepsilon_{i,t})}{d\varepsilon} - \varepsilon_{i,t} \frac{d \log g_j(\varepsilon_{j,t})}{d\varepsilon}] \\ + & \sum_{t=1}^{T} \{ [\varepsilon_{j,t} \frac{d^2 \log g_i(\varepsilon_{i,t})}{d\varepsilon^2} - \frac{d \log g_j(\varepsilon_{j,t})}{d\varepsilon}] \varepsilon_t' \} C_0' \delta_i \\ - & \sum_{t=1}^{T} \{ [\varepsilon_{i,t} \frac{d^2 \log g_j(\varepsilon_{j,t})}{d\varepsilon^2} - \frac{d \log g_i(\varepsilon_{i,t})}{d\varepsilon}] \varepsilon_t' \} C_0' \delta_j = 0, i < j. \end{aligned}$$

Let us now introduce the effect of the number of observations. We get :

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}[\varepsilon_{j,t}\frac{d\log g_i(\varepsilon_{i,t})}{d\varepsilon}-\varepsilon_{i,t}\frac{d\log g_j(\varepsilon_{j,t})}{d\varepsilon}]$$

$$+ \quad E_0\{[\varepsilon_{j,t}\frac{d^2\log g_i(\varepsilon_{i,t})}{d\varepsilon^2}-\frac{d\log g_j(\varepsilon_{j,t}}{d\varepsilon})]\varepsilon_t'\}C_0'\sqrt{T}\delta_i$$

$$- \quad E_0\{\varepsilon_{i,t}\frac{d^2\log g_j(\varepsilon_{j,t})}{d\varepsilon^2}-\frac{d\log g_i(\varepsilon_{i,t})}{d\varepsilon}]\varepsilon_t'\}C_0'\sqrt{T}\delta_j=o_p(1).$$

We have :

i) $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}[\varepsilon_{j,t}\frac{d\log g_i(\varepsilon_{i,t})}{d\varepsilon}-\varepsilon_{i,t}\frac{d\log g_j(\varepsilon_{j,t})}{d\varepsilon}]\underset{d}{\to}Z_{i,j}, i<j,$

where the random vector obtained by stacking the $Z_{i,j}$ is Gaussian with zero-mean and $Cov(Z_{i,j},Z_{k,l})=\Omega_{(i,j),(k,l)}$, where

$$\Omega_{(i,j),(k,l)} = 0, \text{ if } i<j, k<l,$$

$$\Omega_{(i,j),(i,l)} = E_0[\frac{d\log g_j(\varepsilon_{j,t})}{d\varepsilon}]E_0[\frac{d\log g_l(\varepsilon_{l,t})}{d\varepsilon}], \text{ if } j\neq l,$$

$$\Omega_{(i,j),(k,j)} = E_0[\frac{d\log g_i(\varepsilon_{i,t})}{d\varepsilon}]E_0(\frac{d\log g_k(\varepsilon_{k,t})}{d\varepsilon}), \text{ if } i\neq k,$$

$$\Omega_{(i,j),(i,j)} = E_0([\frac{d\log g_i(\varepsilon_{i,t})}{d\varepsilon}]^2)+E_0([\frac{d\log g_j}{d\varepsilon}(\varepsilon_{j,t})]^2),$$

$$- \quad 2E_0[\varepsilon_{i,t}\frac{d\log g_i(\varepsilon_{i,t})}{d\varepsilon}]E_0[\varepsilon_{j,t}\frac{d\log g_j(\varepsilon_{j,t})}{d\varepsilon}],$$

$$\Omega_{(i,j),(k,i)} = -E_0[\frac{d\log g_j(\varepsilon_{j,t})}{d\varepsilon}]E_0[\frac{d\log g_k(\varepsilon_{k,t})}{d\varepsilon}] \text{ (with necessarily } k<j),$$

$$\Omega_{(i,j),(j,l)} = -E_0[\frac{d\log g_i(\varepsilon_{i,t})}{d\varepsilon}]E_0[\frac{d\log g_l(\varepsilon_{l,t})}{d\varepsilon}] \text{ (with necessarily } i<l).$$

ii) Let us now denote :

$$a'_{i,j} = E_0\{-[\varepsilon_{j,t}\frac{d^2 \log g_i(\varepsilon_{i,t})}{d\varepsilon^2} - \frac{d \log g_j(\varepsilon_{j,t})}{d\varepsilon}]\varepsilon'_t\}C'_0$$

$$= \{E_0[-\frac{d^2 \log g_i(\varepsilon_{i,t})}{d\varepsilon^2}] + E_0[\varepsilon_{j,t}\frac{d \log g_j(\varepsilon_{j,t})}{d\varepsilon}]\}c'_{j,0}.$$

Then, $\forall i < j$,

$$a'_{i,j}\sqrt{T}\delta_i - a'_{j,i}\sqrt{T}\delta_j \underset{d}{\to} Z_{i,j}.$$

Let us introduce the notations :

$$\begin{array}{rcl} \delta & = & (\delta'_1, \ldots, \delta'_n)', \delta \text{ is a } n^2 \text{ dimensional vector,} \\ Z & = & (Z_{1,2}, \ldots, Z_{1,n}, Z_{2,3}, \ldots, Z_{2,n}, \ldots, Z_{n-1,n}), \end{array}$$

where Z is a $n(n-1)/2$ dimensional vector

$$A_1 = \begin{bmatrix} a'_{1,2} & -a'_{2,1} & 0 & \ldots & \ldots & 0 & 0 \\ a'_{1,3} & 0 & -a'_{3,1} & & & & \\ a'_{1,n} & \ldots & \ldots & \ldots & \ldots & \ldots & a_{n,1} \\ 0 & a'_{2,3} & -a'_{3,2} & \ldots & & 0 & 0 \\ 0 & a'_{2,4} & 0 & -a'_{4,2} & \ldots & 0 & 0 \\ & \ldots & \ldots & \ldots & \ldots & \ldots & \\ 0 & a'_{2,n} & 0 & \ldots & \ldots & 0 & a'_{n,2} \\ & \ldots & \ldots & \ldots & \ldots & \ldots & \\ 0 & 0 & 0 & \ldots & \ldots & a'_{n-1,n} & a'_{n,n-1} \end{bmatrix},$$

where A_1 is $[n(n-1)/2, n^2]$ matrix,

$$A_2 = \begin{bmatrix} c'_{2,0} & c'_{1,0} & 0 & \dots & \dots & 0 & 0 \\ c'_{3,0} & 0 & c'_{1,0} & \dots & \dots & 0 & 0 \\ & \dots & \dots & \dots & \dots & \dots & \\ c'_{n,0} & \dots & \dots & \dots & \dots & 0 & c'_{1,0} \\ 0 & c'_{3,0} & c'_{2,0} & \dots & \dots & 0 & 0 \\ 0 & c'_{4,0} & 0 & c'_{2,0} & \dots & 0 & 0 \\ & \dots & \dots & \dots & \dots & \dots & \\ 0 & c'_{n,0} & 0 \dots & \dots & \dots & 0 & c'_{2,0} \\ & \dots & \dots & \dots & \dots & \dots & \\ 0 & 0 & 0 & 0 & & c'_{n,0} & c'_{n-1,0} \end{bmatrix},$$

where A_2 is a $[\frac{n(n-1)}{2}, n^2]$ matrix,

$$A_3 = \begin{bmatrix} c'_{1,0} & 0 & \dots & \dots & 0 \\ 0 & c'_{2,0} & \dots & 0 & \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & c'_{n,0} \end{bmatrix},$$

where A_3 is a (n, n^2) matrix.

Then we have :

$$A\sqrt{T}\delta \underset{d}{\to} \begin{pmatrix} Z \\ 0 \end{pmatrix}$$

where $A = \begin{pmatrix} A_1 \\ A_2 \\ A_3 \end{pmatrix}$ is a (n^2, n^2) matrix, or equivalently :

$$\sqrt{T}\delta \underset{d}{\rightarrow} A^{-1} \begin{pmatrix} Z \\ 0 \end{pmatrix}$$

Noting that $\Omega = V(Z)$ is obtained from the terms $\Omega_{(i,j),(k,l)}$ given above, we get the asymptotic distribution of $\sqrt{T}\delta$:

$$\sqrt{T}\delta \approx N[0, A^{-1} \begin{pmatrix} \Omega & 0 \\ 0 & 0 \end{pmatrix} A'^{-1}]$$

which is a Gaussian distribution on a vector subspace of dimension $n(n-1)/2$.

As noted in Pham, Garat (1997), Section 2.B, the first-order expansion of the finite sample FOC depends on $\delta_i = \hat{c}_i - c_{i,0}$ by means of the quantities $c'_{j,0}\delta_i = c'_{j,0}(\hat{c}_i - c_{i0,})$, which are simply the opposite of the elements in the first-order expansion of the contamination coefficients $\hat{\Delta} = Id - C_0^{-1}\hat{C} = Id - C'_0\hat{C}$.

Since $\hat{C} = C_0 + (\delta_1, \ldots, \delta_n)$, we have :

$$\hat{\Delta}_{i,j} = -c'_{i,0}\delta_j.$$

We have the following results :

i) The asymptotic distribution of $\hat{\Delta}$ is degenerate, since

$$\sqrt{T}(\hat{\Delta}_{i,j} + \hat{\Delta}_{j,i}) = o_p(1), i < j,$$

$$\sqrt{T}(\hat{\Delta}_{i,i}) = o_p(1), i = 1, \ldots, n,$$

due to the expansion of the conditions for the orthogonal matrix $\hat{C}$.

ii) Thus the asymptotic distribution of $\hat{\Delta}$ is known whenever we know the asymptotic distribution of its strictly lower triangular part, that is, of the $\hat{\Delta}_{i,j}, i < j$.

iii) The joint distribution of the $\hat{\Delta}_{i,j}, i < j$, is easily deduced by using the definition of $a_{i,j}$ and the convergence :

$$a'_{i,j}\sqrt{T}\delta_i - a'_{j,i}\sqrt{T}\delta_j \underset{d}{\rightarrow} Z_{i,j}.$$

We get :

$$\sqrt{T}E_0[\frac{d^2\log g_i(\varepsilon_{i,t})}{d\varepsilon^2}+\frac{d^2\log g_j(\varepsilon_{j,t})}{d\varepsilon^2}-\varepsilon_{j,t}\frac{d\log g_j(\varepsilon_{j,t})}{d\varepsilon}-\varepsilon_{i,t}\frac{d\log g_i(\varepsilon_{i,t})}{d\varepsilon}]\hat{\Delta}_{i,j} \underset{d}{\to} Z_{i,j}.$$

The factor multiplying $\hat{\Delta}_{i,j}$ is nonzero, because of the local concavity condition, and the asymptotic distribution of the $\hat{\Delta}_{i,j}, i<j$, is derived.

As in Pham, Garat (1997), the asymptotic distribution of the $\hat{\Delta}_{i,j}$ no longer depends on matrix C, but just on the distributional properties of the sources and on the choice of the pseudo p.d.f.

Our results have taken explicitly into account the constraints of orthogonal matrix C in the first-order conditions. In this respect our expansions differ from the expansions in Pham, Garat (1997) or Wei (2014) as well as the associated asymptotic distribution of the estimators.

A.4.2. Recursive PML Estimator

The FOC of the finite sample optimization problem (3.8) are :

$$\begin{cases} \sum_{t=1}^{T} Y_t \frac{d\log g_i}{d\varepsilon}(\hat{c}_i' Y_t) - \sum_{j=1}^{i} \hat{\lambda}_{i,j}\hat{c}_j = 0, i=1,\ldots,n, \\ \\ \hat{c}_i'\hat{c}_j = 0, j<i, \hat{c}_i'\hat{c}_i = 1, i=1,\ldots,n, \end{cases}$$

where $\hat{\lambda}_{i,j}, j<i$ (resp. $\hat{\lambda}_{i,i}$) is the estimated Lagrange multiplier associated with the restriction $c_i'\hat{c}_j = 0, j<i$ (resp. $c_i'c_i = 1$).

Note that at the n^{th} iteration $\hat{c}_n$ is (essentially) characterized by the orthogonality restrictions.

As for deriving system (2.9) of FOC for the PML estimator, we can premultiply the first subsystem by $\hat{C}'$. We get :

$$\sum_{t=1}^{T} \hat{c}_j' Y_t \frac{d\log g_i}{d\varepsilon}(\hat{c}_i' Y_t) - \hat{\lambda}_{i,j} = 0, j \leq i.$$

Then we can substitute this expression of the Lagrange multiplier in the system to get :

$$\sum_{t=1}^{T}[Y_t\frac{d\log g_i}{d\varepsilon}(\hat{c}_i'Y_t)-\sum_{j=1}^{i}(\hat{c}_j'Y_t\frac{d\log g_i}{d\varepsilon}(\hat{c}_i'Y_t)\hat{c}_j)]=0, i=1,\ldots,n,$$

$$\iff \sum_{t=1}^{T}\{\frac{d\log g_i}{d\varepsilon}(\hat{c}_iY_t)[Y_t-\sum_{j=1}^{i}\hat{c}_j'Y_t\hat{c}_j]\}=0, i=1,\ldots,n.$$

This system is easily solved recursively.

Appendix 5

Asymptotic Variance of the PML Estimator for $n = 2$.

A.5.1. Derivation of the asymptotic variance

When $n = 2$, the orthogonal matrix C (with $\det C = 1$) can be parametrized as :

$C(\theta) = \begin{pmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{pmatrix}$ and the pseudo log-likelihood function written as :

$$L_T(\theta) = \sum_{t=1}^{T} \{\log g_1[c_1'(\theta)y_t] + \log g_2[c_2'(\theta)y_t]\} \equiv \sum_{t=1}^{T} \log f(y_t;\theta).$$

The PML estimator of parameter θ is asymptotically normal with variance.

$$V_{as}[\sqrt{T}(\hat{\theta}_T - \theta_0)] = J^{-2}I,$$

where $J = E_0\left[\frac{-\partial^2 \log f(Y_t;\theta_0)}{\partial\theta^2}\right], I = E_0\left(\left[\frac{\partial \log f(Y_t;\theta_0)}{\partial\theta}\right]^2\right).$

We have :

$$\begin{aligned}
\frac{\partial \log f(y_t;\theta)}{\partial\theta} &= \sum_{i=1}^{2}\left\{\frac{d\log g_i}{d\varepsilon}[c_i'(\theta)y_t]\frac{dc_i'(\theta)}{d\theta}y_t\right\}, \\
\frac{\partial^2 \log f(y_t;\theta)}{\partial\theta^2} &= \sum_{i=1}^{2}\left\{\frac{d\log g_i}{d\varepsilon}[c_i'(\theta)y_t]\frac{d^2c_i'(\theta)}{d\theta^2}y_t\right. \\
&+ \left.\frac{d^2\log g_i}{d\varepsilon^2}[c_i'(\theta)y_t]\left[\frac{dc_i'(\theta)}{d\theta}y_t\right]^2\right\}.
\end{aligned}$$

It is easily checked that :

$$\frac{dC'(\theta)}{d\theta}C(\theta) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \frac{d^2C'(\theta)}{d\theta^2}C(\theta) = -Id.$$

We deduce that :

$$\begin{aligned}
\frac{\partial \log f(y_t;\theta_0)}{\partial\theta} &= \frac{d\log g_1(\varepsilon_{1,t})}{d\varepsilon}\varepsilon_{2,t} - \frac{d\log g_2(\varepsilon_{2,t})}{d\varepsilon}\varepsilon_{1,t},\\
\frac{\partial^2 \log f(y_t;\theta_0)}{\partial\theta^2} &= -\frac{d\log g_1(\varepsilon_{1,t})}{d\varepsilon}\varepsilon_{1,t} - \frac{d\log g_2(\varepsilon_{2,t})}{d\varepsilon}\varepsilon_{2,t}\\
&+ \frac{d^2\log g_1(\varepsilon_{1,t})}{d\varepsilon^2}\varepsilon_{2,t}^2 + \frac{d^2\log g_2(\varepsilon_{2,t})}{d\varepsilon^2}\varepsilon_{1,t}^2.
\end{aligned}$$

Thus :

$$\begin{aligned}
I &= E_0\left[\left(\frac{d\log g_1(\varepsilon_{1,t})}{d\varepsilon}\right)^2\right] + E_0\left[\left(\frac{d\log g_2(\varepsilon_{2,t})}{d\varepsilon}\right)^2\right]\\
&\quad -2E_0\left[\varepsilon_{1,t}\frac{d\log g_1(\varepsilon_{1,t})}{d\varepsilon}\right]E_0\left[\varepsilon_{2,t}\frac{d\log g_2(\varepsilon_{2,t})}{d\varepsilon}\right],\\
J &= E_0\left[\varepsilon_{1,t}\frac{d\log g_1(\varepsilon_{1,t})}{d\varepsilon}\right] + E_0\left[\varepsilon_{2,t}\frac{d\log g_2(\varepsilon_{2,t})}{d\varepsilon}\right]\\
&\quad -E_0\left[\frac{d^2\log g_1(\varepsilon_{1,t})}{d\varepsilon^2}\right] - E_0\left[\frac{d^2\log g_2(\varepsilon_{2,t})}{d\varepsilon^2}\right].
\end{aligned}$$

The asymptotic distribution of $\hat{C} = C(\hat{\theta})$ is deduced by the δ-method, noting that :

$$\frac{dC(\theta)}{d\theta} = \begin{pmatrix} -sin\theta & -cos\theta \\ cos\theta & -sin\theta \end{pmatrix} = [c_2(\theta), -c_1(\theta)].$$

We get :

$$\begin{aligned}
&V_{as}[\sqrt{T}(vec\hat{C} - vecC_0)]\\
&= I/J^2\ vec\left(\frac{dC(\theta_0)}{d\theta}\right)vec\left(\frac{dC(\theta_0)}{d\theta}\right)'.\\
&= I/J^2\begin{bmatrix} c_2(\theta_0)c_2'(\theta_0) & -c_2(\theta_0)c_1'(\theta_0) \\ -c_1(\theta_0)c_2'(\theta_0) & c_1(\theta_0)c_1'(\theta_0) \end{bmatrix}.
\end{aligned}$$

Let us finally discuss the expressions of I and J, when $g_i = f_{i,0}$ is the true distribution. We can construct different parametric models from distribution f_0, that are :

a model with drift parameter $f_0(\varepsilon - m)$;

a model with scale parameter $cf_0(c\varepsilon)$.

From the model with drift parameter, we deduce :

$$E_m\left[\left(\frac{\partial \log f_0(\varepsilon - m)}{\partial m}\right)^2\right] = E\left[-\frac{\partial^2 \log f_0(\varepsilon - m)}{\partial m^2}\right],$$

which for $m = 0$ implies.

$$E_0\left[\frac{d \log f_0(\varepsilon)}{d\varepsilon}\right]^2 = E_0\left[\frac{-d^2 \log f_0(\varepsilon)}{d\varepsilon^2}\right].$$

From the model with scale parameter, we deduce a zero-mean score :

$$E_c\left[\frac{1}{c} + \varepsilon\frac{d \log f_0(c\varepsilon)}{d\varepsilon}\right] = 0,$$

which implies for $c = 1$:

$$E_0\left[\varepsilon\frac{d \log f_0(\varepsilon)}{d\varepsilon}\right] = -1.$$

Thus, if $g_i = f_{i,0}, i = 1, 2$, we get as expected the same value for I and J :

$$I = J = \sum_{i=1}^{2} E_0\left[-\frac{d^2 \log f_{i,0}(\varepsilon_{i,t})}{d\varepsilon^2} - 1\right].$$

A.5.2. Asymptotic variance of the contamination coefficients.

Let us denote c^1, c^2 the rows of matrix C^{-1}. We have :

$$c^j c_j = 1, j = 1, 2, c^j c_i = 0, \text{ if } i \neq j.$$

With these notations, we get :

$$C^{-1}\hat{C} = \begin{pmatrix} c^1 \\ c^2 \end{pmatrix} (\hat{c}_1, \hat{c}_2) = \begin{pmatrix} c^1\hat{c}_1 & c^1\hat{c}_2 \\ c^2\hat{c}_1 & c^2\hat{c}_2 \end{pmatrix},$$

and $vec(C^{-1}\hat{c}) = (c^1\hat{c}_1, c^2\hat{c}_1, c^1\hat{c}_2, c^2\hat{c}_2)'$.

The elements of the asymptotic variance $vec\hat{\Delta}$ are equal to the elements of the asymptotic variance of $vec(C^{-1}\hat{c})$. They are easily computed. For instance we have :

$$\begin{aligned} & V_{as}[\sqrt{T}(c^1\hat{c}_1 - 1)] \\ & = \frac{\omega^2}{(\gamma_{1,2} + \gamma_{2,1})^2} c^1 c_2 c_2' (c^1)' = 0, \\ & V_{as}[\sqrt{T}c^2\hat{c}_1] = \frac{\omega^2}{(\gamma_{1,2} + \gamma_{2,1})^2} c^2 c_2 c_2' = \frac{\omega^2}{(\gamma_{1,2} + \gamma_{2,1})^2} \end{aligned} \quad ,$$

and so on.

Appendix 6

Expansion of the Empirical Covariance

Let us consider i.i.d. observations $(X_t, Y_t), t = 1, \ldots, T$. Their empirical covariance can be expanded for large T as :

$$\begin{aligned}
& \sqrt{T}[\widehat{Cov}(X,Y) - Cov(X,Y)] \\
= \;& \sqrt{T}\{\frac{1}{T}\sum_{t=1}^{T}[X_tY_t - E(XY)] - \frac{1}{T}\sum_{t=1}^{T}X_t\frac{1}{T}\sum_{t=1}^{T}Y_t + EXEY\} \\
\simeq \;& \sqrt{T}\{\frac{1}{T}\sum_{t=1}^{T}[X_tY_t - E(XY)] - \frac{1}{T}\sum_{t=1}^{T}(X_t - EX)EY - \frac{1}{T}\sum_{t=1}^{T}(Y_t - EY)EX\} + o_P(1) \\
= \;& \frac{1}{\sqrt{T}}\sum_{t=1}^{T}[(X_t - EX)(Y_t - EY) - Cov(X,Y)] + o_P(1).
\end{aligned}$$

This expansion can be used to compute the asymptotic variance of an empirical covariance as well as the asymptotic covariance between two empirical covariances. For instance we have :

$$V_{as}[\sqrt{T}[\widehat{Cov}(X,Y) - Cov(X,Y)]] = V[(X - EX)(Y - EY)],$$

$$Cov_{as}\{\sqrt{T}[\widehat{Cov}(X,Y) - Cov(X,Y)|, \sqrt{T}[\widehat{Cov}(Z,U) - Cov(Z,U)]\}$$

$$= Cov[(X - EY)(Y - EY), (Z - EZ)(U - EU)].$$

Appendix 7

Empirical Likelihood Approach

The empirical likelihood approach has been suggested to reach the semi-parametric efficiency bound in IC models [Bach, Jordan (2002), p7]. This appraoch is based on the minimization of a Kullback-Leibler Information Criterion (KLIC).

It is known that the KLIC is a contrast between probability distributions, that it is not symmetric, and that one of its definitions is more appropriate for the optimization. More precisely it has been shown in Kitamura et al. (2004) that, when the parameter of interest is defined by moment restrictions, the empirical likelihood estimator coincides with an efficient GMM estimator. Let us check if this property is still valid for an IC model (with covariance restrictions). For expository purpose we consider the bidimensional case.

Let us denote by $h_0(y_t)$ the true joint density function of $Y_t = (Y_{1,t}, Y_{2,t})'$ and by $h(y_t; \alpha, g_1, g_2) = g_1[c_1'(\alpha)y_t]g_2[c_2'(\alpha)y_t]$, the restricted density corresponding to the semi-parametric IC model. The latter now depends on parameter α characterizing matrix C, but also on functional parameters g_1, g_2 corresponding to the densities of the sources. The asymptotic optimization of the KLIC is :

$$\min_{\theta, g_1, g_2} \int\int h(y; \alpha, g_1, g_2) \log \frac{h_0(y)}{h(y; \alpha, g_1, g_2)} dy, \tag{a.1}$$

$$s.t. \int g_1(\varepsilon_1)d\varepsilon_1 = 1, \int \varepsilon_1 g_1(\varepsilon_1)d\varepsilon_1 = 0, \int \varepsilon_1^2 g_1(\varepsilon_1)d\varepsilon_1 = 1,$$

$$\int g_2(\varepsilon_2)d\varepsilon_2 = 1, \int \varepsilon_2 g_2(\varepsilon_2)d\varepsilon_2 = 0, \int \varepsilon_2^2 g_2(\varepsilon_2)d\varepsilon_2 = 1,$$

where the constraints[13] correspond to the assumptions $E\varepsilon_t = 0, V\varepsilon_t = Id$.

In finite sample the optimization problem is of the same type with the finite sample objective function :

[13]Note that the constraint $\int\int \varepsilon_1\varepsilon_2 g_1(\varepsilon_1)g_2(\varepsilon_2)d\varepsilon_1 d\varepsilon_2 = 0$ is automatically satisfied.

$$\int\int h(y;\alpha,g_1,g_2)\log\frac{\hat{h}_T(y)}{h(y;\theta,g_1,g_2)}dy, \tag{a.2}$$

where $\hat{h}_T$ is for instance a kernel estimator of the joint density of $(Y_{1,t},Y_{2,t})$.

Let us now examine the FOC, with a focus on the FOC corresponding to functions g_1,g_2. Indeed, by analogy with Kitamura et al. (2004), we have to check if it is possible to solve explicitly these FOC with respect to g_1,g_2, and then to derive the concentrated KLIC, function of α only. By a change of variable, the objective function in the optimization problem is equivalently written as :

$$\int\int g_1(\varepsilon_1)g_2(\varepsilon_2)\log\left[\frac{\hat{h}_T(c_1(\alpha)\varepsilon_1+c_2(\alpha)\varepsilon_2)}{g_1(\varepsilon_1)g_2(\varepsilon_2)}\right]d\varepsilon_1 d\varepsilon_2.$$

Let us denote by $\lambda_{i,j}, i=1,2, j=0,1,2$, the Lagrange multipliers associated with the constraints in (a.1) and optimize the Lagrangian with respect to the pseudo p.d.f. g_1. The FOC of the variational problem for g_1 are :

$$\int g_2(\varepsilon_2)[\log\frac{\hat{h}_T(c_1(\theta)\varepsilon_1+c_2(\theta)\varepsilon_2)}{g_1(\varepsilon_1)g_2(\varepsilon_2)}-1]d\varepsilon_2-\lambda_{1,0}-\lambda_{1,1}\varepsilon_1-\lambda_{1,2}\varepsilon_1^2=0,\forall\varepsilon_1.$$

These FOC become :

$$-\log g_1(\varepsilon_1)+\int g_2(\varepsilon_2)\log\frac{\hat{h}_T(c_1(\alpha)\varepsilon_1+c_2(\alpha)\varepsilon_2)}{g_2(\varepsilon_2)}d\varepsilon_2-(\lambda_{1,0}+1)-\lambda_{1,1}\varepsilon_1-\lambda_{1,2}\varepsilon_1^2=0,\forall\varepsilon_1,$$

and similar conditions for the optimization w.r.t. g_2 :

$$-\log g_2(\varepsilon_2)+\int g_1(\varepsilon_1)\log\frac{\hat{h}_T(c_1(\alpha)\varepsilon_1+c_2(\alpha)\varepsilon_2)}{g_1(\varepsilon_1)}d\varepsilon_1-(\lambda_{2,0}+1)-\lambda_{2,1}\varepsilon_2-\lambda_{2,2}\varepsilon_2^2=0,\forall\varepsilon_2.$$

In the standard case of moment restrictions [Kitamura et al. (2004)], the FOC do not contain the integral terms. They can be solved up to the multipliers, and then the multipliers deduced from the constraints in (a.1). In the case of an IC model, these equations can be solved numerically, but do not provide a closed form expression for the solution in g_1 and g_2.

Appendix 8

A Review of Semi-Parametrically Efficient Methods

The semi-paremetrically efficient methods try to estimate jointly the mixing matrix and the distribution of the errors. For instance such an approach is followed in Samarov, Tsybakov (2004), who base the estimation on the restricted p.d.f. of the observable and derive estimator of the mixing matrix by considering the spectral decomposition of the matrix $E_0[\frac{df(Y)}{dy}\frac{df(Y)}{dy'}]$ approximated by kernel, where $f(y)$ is the p.d.f. of Y. More generally a joint spectral decomposition can be performed on two other scatter matrices, such as the scatter matrices constructed from the second and fourth-order moments as in the fourth-order blind identification (FOBI) method [see e.g. Cardoso (1989), Oja et al. (2006), Bonhomme, Robin (2009), Ilmonen (2015)]. Alternatively Chen, Bickel (2005) consider the restricted characteristic function of the errors and minimize a distance between the constrained and unconstrained characteristic functions. Statistical properties of these estimation methods are derived such as the convergence in Samarov, Tsybakov (2004), the convergence and the asymptotic distribution in Chen, Bickel (2005) or Ilmonen (2015). These approaches do not reach the semi-parametric efficiency bound.

Recent papers have introduced more complex procedures to compute the semi-parametric efficiency bound and try to reach this bound. For instance Ilmonen, Paindavene (2015) consider the special case of an IC model, where the errors have symmetric distributions with common median zero and derive the associated semi-parametric efficiency bound. By using the fact that the maximal invariant of this model is the vector of marginal signed ranks of the residuals, they construct an efficient estimator by inverting the test statistics of the null hypotheses [14] $H_0 : (C = C_0)$, **with given densities**. They show that these estimators are consistent even with misspecified densities (the analogue of the consistency result for PML estimator), but they do not achieve the efficiency under this misspecification .

Chen, Bickel (2006) propose to estimate the mixing matrix by solving the condition of zero efficient score, after substitution of an estimated score to

[14] However their asymptotic distribution has not yet been derived.

the true one. This approach requires a first step consistent, but not necessarily efficient estimator, to approximate the score. This first step estimator can be a PML estimator, or a generalized covariance estimator. Alternative approaches approximate directly the likelihood function by assuming the densities of the sources in large parametric families, such as mixtures of Gaussian distributions [Vlassis (2001), Eloyan, Ghosh (2011)], or exponentially spline tilted Gaussian densities [Hastie, Tibshirani (2002)]. In such methods the estimation of the efficient score is updated at each step of the optimization algorithm, either an EM algorithm, or a Newton-Raphson algorithm. Other approaches are likely efficient, such as the two step approach described in the second remark after Corollary 1, or the empirical likelihood approach (see Appendix 7). However their asymptotic distributions have not yet been derived.